EUI DEPARTMENT OF ECONOMICS

Microeconometrics Applied to Labour and Migration

Damiano Argan

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Florence, 19 February 2024

European University Institute
Department of Economics

Microeconometrics Applied to Labour and Migration

Damiano Argan

Thesis submitted for assessment with a view to
obtaining the degree of Doctor of Economics
of the European University Institute

Examining Board

Prof. Andrea Mattozzi, University of Bologna and EUI, Supervisor
Prof. Andrea Ichino, EUI
Prof. Robert Gary-Bobo, Universite Paris I Pantheon-Sorbonne
Prof. Fabiano Schivardi, LUISS

Researcher declaration to accompany the submission of written work
Department Economics - Doctoral Programme

I <Damiano Argan> certify that I am the author of the work <Microeconometrics Applied to Labour and Migration> I have presented for examination for the Ph.D. at the European University Institute. I also certify that this is solely my own original work, other than where I have clearly indicated, in this declaration and in the thesis, that it is the work of others.

I warrant that I have obtained all the permissions required for using any material from other copyrighted publications.

I certify that this work complies with the Code of Ethics in Academic Research issued by the European University Institute (IUE 332/2/10 (CA 297).

I declare that this work consists of <53937> words.

Signature and date:

13/12/2023
Damiano Argan

A Mamma, Papà e Giovanni,
E agli amici di una vita, che son come fratelli.

Abstract

The thesis contains four independent essays that focus on economics of migration, education and labour.

The first chapter joint with Anatole Cheysson focus on the effect of plurilingualism on brain drain. We study how foreign language proficiency affects brain drain by exploiting the heterogenous exposure of Albania to Italian television in the second half of the twentieth century. We document that Albanians' exposure to the Italian TV signal was as good as random. We find that exposure to Italian TV led to a considerable increase in Italian proficiency rates and strongly increased the probability of migrating of highly skilled individuals while not affecting other skill groups.

The second chapter, joint with Robert Gary-Bobo and Marion Goussé, studies the variation of the average treatment effect of education over time. To study the returns to degrees, we assume the existence of a finite number of latent types and estimate a finite-mixture model. We show that the expected real wages commanded by some higher-education degrees decreased in absolute terms in France, in the past two decades, and that this drop is not due to adverse selection. In the case of Master degrees, the student selection improved with time, in spite of the fact that the number of graduates increased substantially.

The third chapter critically reviews the literature on the human capital of entrepreneurs exploiting insights derived from previous studies and stylized facts obtained from high-quality data in Denmark's administrative register.

The fourth chapter, co-authored with Leonardo Indraccolo and Jacek Piosik, studies the human capital determinants of entrepreneurship. Using Danish administrative data, we measure analytical and communication skills with high school grades in math and Danish language. We observe a positive complementarity between math and Danish language grades in predicting individuals' self-selection

into entrepreneurship. For the population of high performing math students, we exploit within-school, across-cohort variation in students' exposure to peers whose father has a university degree in humanities to identify the effect of communication skills on the probability of becoming an entrepreneur. We find that the difference in entrepreneurship share between the most (90th percentile) and the least (10th percentile) exposed individual is 1.1 percentage points: 20% of the overall share of entrepreneurs in the economy.

Acknowledgements

I thank my supervisors Andrea Mattozzi, Andrea Ichino, and Robert Gary-Bobo for having taught me economics with great patience and dedication. I will owe them all my life, as the gift of knowledge is everlasting.

I thank Andrea Mattozzi for his nurturing guidance that encompasses economics, how to write a paper and complete a thesis, and above all, for being a reassuring presence I have always known would be by my side, no matter what happened.

I express my gratitude to Andrea Ichino for his exceptional teaching of econometrics and causality. I believe that I will owe him for shaping my career and providing me with the skills to sustain my life.

I thank Robert Gary-Bobo for his paternal guidance that has endured for seven years. He taught me the fundamentals of the profession, and I owe him for inspiring my decision to pursue my PhD at EUI, something for which I will be forever thankful.

Certainly, my PhD thesis would have never seen the light, and most of my accomplishments would not have been possible without my family. I thank my Mom for dedicating her entire life to her sons with love, dedication, and sacrifice. There hasn't been a single day when I couldn't turn to her in times of need, and there hasn't been a day when I didn't feel her love.

I thank my Dad for supporting me throughout my life with his time, dedication, and insightful discussions. His example of how a person should be, both in his profession and in his interactions with others, has consistently motivated me to become the best version of myself.

I thank my brother a life companion for which we do not know nothing but love and solidarity for each other.

Lastly, I express my gratitude to Maria. If she were here, I am certain she would have been incredibly supportive and joyful about my achievements. Making her proud would have brought me immense happiness.

I thank Anatole, Hugo, Leonardo, Carlita and Anna the companions of this journey, and this fight, without whom I would not have laugh, had fun, and felt that it was more than simply write a PhD thesis but an experience that made me new life companions and soulmates.
I thank the friends of my entire life, whom I can only list in alphabetical order, as we are all equal to one another. Here are their names: Andreana, Augu, Carlo, Daniel, Edoa, Gianvi, Lollo, Matteo, Michi, Michu, Osvi, Peter, Pucio, Ric, Stroffo, Tullio.
Lastly, I want to convey my gratitude to Sarah for the time we've spent together and her patience with me. I also extend my thanks to Antonella and Loredana for their maternal care, the shared laughter, and the mutual support we've had over the past five years. Lastly, I'm thankful to Maurizio for the days we've spent together at the institute. I will miss all of you.

Contents

Chapter 1

Plurilingualism and Brain Drain: Unexpected Consequences of Access to Foreign TV

Co-written with **Anatole Jacques Idrissa Cheysson**

1.1 Abstract

We study how foreign language proficiency affects brain drain by exploiting the heterogenous exposure of Albania to Italian television in the second half of the twentieth century. We document that, due to geographical proximity, the Italian TV signal accidentally reached Albania and, conditional on geographic conditions, Albanians' exposure to the signal was as good as random. We find that exposure to Italian TV led to a considerable increase in Italian proficiency rates and strongly increased the probability of migrating of highly skilled individuals while not affecting other skill groups.

1.2 Introduction

Linguistic distance between countries' languages is a key determinant of migratory flows (Belot and Ederveen, 2012; Adserà and Pytliková, 2015). As this distance increases, migrants tend to experience poorer labor market results (Adsera and Ferrer, 2015), which in turn makes migrating to countries with significantly different languages less appealing. The penalty imposed by linguistic differences is especially hard on high-skill individuals for whom communication skills are more valuable (Chiswick, 1995; Berman et al., 2003). Consequently, linguistic distance is an important driver of migrants' self-selection into emigration (Borjas, 1987; Be-

lot and Hatton, 2012): the higher the proximity between two countries' languages, the more migratory flows are composed of high skilled individuals.

Although the relationship between language and migration has received much attention, empirical research has remained observational in nature, unable to quantify and inform the causal effect of foreign language proficiency on migration decisions. To this day, we have little evidence on the effects of policies that promote plurilingualism on emigration patterns, and in particular on the emigration decisions of the educated. However, the considerable impact of the migration of high-skill individuals, i.e. brain drain, on the economy of origin countries is the subject of ongoing interest in the literature (Docquier and Rapoport, 2012; Shrestha, 2017; Anelli et al., 2023). In a related context, EU policy makers are keen on evaluating the impact of language barriers on labor mobility, which is crucial for the success of monetary unions.[1] We fill this gap in the literature by providing novel causal evidence on the relationship between foreign language proficiency and emigration, with a specific focus on the emigration of high-skill individuals.

In 1957, the Italian public broadcasting company (RAI) built a TV transmitter in Puglia, a region in southeast Italy, and its signal inadvertently reached parts of neighboring Albania. During that period, and until 1990, Albania was a communist dictatorship isolated from the rest of the world, both physically and culturally. Conditional on geographical characteristics, we show that individual's exposure to Italian television was quasi-random; the specificity of this historical episode dispels common endogeneity concerns as signal access was unintentional and internal movement in Albania was restricted, preventing Albanians from relocating to areas with signal availability. These factors address the problem of endogenous location choice for both the transmitter and individuals, which typically results in biased estimates of media exposure on observed outcomes. Following the collapse of the regime in 1990, massive emigration waves and a brain drain occurred (Gërmenji and Milo, 2011; Gëdeshi and King, 2019). We leverage this distinctive context to examine the impact of Italian television access on Italian language proficiency and the Albanian brain drain.

For this study, we use three datasets : (i) a geo-referenced dataset of signal

[1] Since Mundell (1961), mobility has been considered key to the success of monetary unions. In the EU, labor is deemed not mobile enough, especially when compared to the US labor market (House et al. (2018)). Language barriers are seen by EU policy makers as one of the reasons for the lack of labor mobility (https://education.ec.europa.eu/focus-topics/improving-quality/multilingualism/about-multilingualism-policy, among others).

availability and power provided by RAI (*RAI* dataset, henceforth); (ii) a geo-referenced dataset of terrain characteristics aggregated at the municipality level (*Geographic* dataset); (iii) the 2005 Living Standard Measurement Survey conducted by the World Bank and Albanian statistical agency (*LSMS* dataset). We measure the average exposure to Italian television for each municipality in Albania using the *RAI* dataset. With the *Geographic* dataset, we generate a comprehensive set of geographical and topographic controls at the municipality level. We exploit three sections of the LSMS: the internal migration folder to relocate individuals to their municipality of residence in 1990 to infer their access to Italian television prior to the dictatorship's fall; the 1990 foreign language proficiency questionnaire; and since the LSMS, by design, only includes individuals residing in Albania in 2005, we use the questionnaire on respondents' siblings' residences to create a dataset that encompasses migrants.

Our study offers two novel contributions. Firstly, we examine the impact of Italian television access on foreign language proficiency in 1990. Since the LSMS base sample consists only of non-migrants, we cannot estimate the average treatment effect of television on language proficiency. However, in line with the literature, we assume that non-migrants have a lower propensity to learn a foreign language than migrants (Bütikofer and Peri, 2021), implying that estimating the effect of Italian TV access on language skills for non-migrants provides a lower bound. We estimate a lower-bound positive increase of 7 percentage points in Italian proficiency rates between municipalities fully exposed to Italian television and those with no exposure, which is more than double the average Italian language proficiency rate of 5.3% in 1990. We also successfully conduct placebo tests for other foreign languages.

Our second contribution involves estimating the causal impact of Italian TV exposure on the likelihood of emigration. Using the sample of LSMS respondents' siblings, we observe no effect on the probability of migration when estimating for the entire sample. However, for high-skilled individuals, we find a substantial positive effect of approximately 20 percentage points on the likelihood of emigrating abroad, accompanied by a similar effect on the probability of emigrating to Italy. While we cannot estimate an instrumental variable regression to extend beyond this reduced form estimate due to data limitations, the already sizable 20 percentage point increase allows us to confidently assert that foreign language proficiency significantly boosts the migration probability for high-skilled individuals.

We then discuss the exclusion restrictions, specifically that exposure to Ital-

ian TV only influences migration behavior through Italian language knowledge. Competing channels include television's role as an information provider and its impact on expected returns from migration (Farré and Fasani, 2013; Pesando et al., 2021; Adema et al., 2022). We exploit interviews conducted in 1991 with Albanian migrants, which reveal that their primary viewing preferences were entertainment programs that lacked pertinent migration information, such as job opportunities, regional economic conditions, mobility, and housing-related details. Furthermore, we use the LSMS to show that Albanians who migrated abroad and returned did not use TV as an information source to organize their emigration. Lastly, we discuss whether Italian TV led Albanians to overestimate the benefits of migrating to Italy (Mai, 2004). However, such a channel would imply a uniform impact across skill categories, which is not what we observed - we only found an effect among high-skilled individuals. This finding corresponds with the notion that language proficiency is crucial for high-skilled migrants, as effective communication skills are particularly important in high-skill jobs, as identified in the literature and predicted by the Borjas model (Borjas, 1987; Chiswick, 1995; Berman et al., 2003).[2]

The paper is organized as follows: Section 1.3 presents the literature review, in Section 1.4 we summarise the historical background, Section 1.5 describes the data, Section 1.6 then discusses our identification strategy, in section 1.7 we show the results. Section 1.8 discusses the exclusion restriction, Section 1.9 presents robustness tests. Finally, Section 1.10 concludes.

1.3 Literature Review

Our paper makes contributions to five areas of literature. Firstly, this study adds to the literature on linguistic and cultural determinants of migration by providing the first causal evidence of language proficiency's effect on emigration patterns. Specifically, we contribute to the literature on the causes of brain drain, which has identified cultural distance as a key predictor of migrants' educational selectivity (Belot and Hatton, 2012). While existing literature has been observational (Belot and Ederveen, 2012; Adsera and Ferrer, 2015), our research presents causal findings that can inform policymakers about the consequences of foreign media exposure and policies promoting plurilingualism.

[2] In the Online Appendix, we show that the Borjas model predicts an increase in migration probabilities for above-average productive individuals as a consequence of a positive exogenous shock to the correlation coefficients for a wide range of parameters.

Second, our study is linked to the literature on the influence of mass media on societal outcomes, from which we derive our identification approach (Olken, 2009; La Ferrara, 2016; Durante et al., 2019).[3] In particular, Farré and Fasani (2013) shows how TV exposure in rural Indonesia reduced internal migration by helping to correct overestimated returns to internal mobility; Adema et al. (2022) shows how internet access increases desire to migrate and actual migration by reducing the cost of information, trust in government and perceived well-being.[4] Our findings complement this research in providing evidence of the effect of media exposure on migration through language skill acquisition, a specific form of human capital, as opposed to other types of information.

Additionally, we connect to research that concentrates on the media's influence on educational outcomes: Gentzkow and Shapiro (2008) shows how television exposure in the US had a positive effect on the test scores of children raised in non-English speaking households. Kearney and Levine (2019) shows that the edutainment program *Sesame Street* was beneficial for children's educational attainment. Durante et al. (2019) demonstrates how children exposed to Berlusconi's television became less cognitively sophisticated and civically minded. Our research complements these findings by showing how exposure to foreign media increased foreign language proficiency.

Finally, this article relates to the research on language proficiency and migrants integration. Causal studies have documented how proficiency in the host country's language increases migrants earnings (Sarvimäki and Hämäläinen, 2016), labour force participation (Lochmann et al., 2019) and employment (Lang, 2022; Schmid, forthcoming). Given our findings that foreign language proficiency increases emigration, our work suggests that potential migrants anticipate these improved labor market outcomes.

[3] This literature includes a wide range of possible outcomes: political outcomes (Gentzkow and Shapiro, 2008; Olken, 2009; Enikolopov et al., 2011), gender norms (Jensen and Oster, 2009; Chong and La Ferrara, 2009; Ferrara et al., 2012; Kearney and Levine, 2015)), and consumption choices (Bursztyn and Cantoni, 2016).

[4] In a 2007 working paper, Braga (2007) explores the influence of Italian television on promoting seasonal migration from Albania. However, the study has notable limitations: it fails to suggest a mechanism through which TV impacts migration, neglects to investigate the role of Italian television in international migration, and does not address the varying effects of TV on different skill groups.

1.4 Historical Background

Enver Hoxha came to power in Albania in 1944 in the immediate aftermath of the war.[5] He rapidly seized absolute power and organized the complete isolation of the country from the outside world: internal migration was controlled and limited, and emigration to foreign countries was forbidden.[6] This isolation also extended to culture: no foreign books, movies, nor newspapers were allowed to circulate. Hoxha's communist regime lasted until 1990.

Despite Enver Hoxha's best efforts, there was *a tear in the wall*. In 1957, the RAI (Radiotelevisione italiana - Italian State Television) built a television transmitter in Martina Franca (Italy, Puglia- the Italian region closest to Albania, on the other side of the sea). Thanks to its power and the short distance between Italy and Albania, the transmitter unintentionally reached parts of Albania, it still broadcasts to this day, and did so without interruption since 1957. Since the 70s, when TV sets began to be widespread in Albanian homes, Albanians have regularly watched Italian television.[7] Italian programs provided entertainment shows that Albanian television did not feature at the time: it only had one channel broadcasting four hours each day, alternating between propaganda and few Albanians films repeated continuously. It is the entertainment content of Italian programming that proved attractive to Albanians.[8]

In 1990, following pressure for reform from the population, the communist structures began to be dismantled, and in 1992 the first democratically elected government took power. From June 1990 onward, Albanians recovered their ability to emigrate. During the 1990s decade, around 800 thousands Albanians migrated abroad, about one fourth of the entire Albanian population at the time. It is estimated that about 600,000 Albanians emigrated to Greece and 200,000 to Italy.[9] This emigration wave has been coined repeatedly as a brain drain in the literature:

[5] This section owes much to Dorfles and Gatteschi (1991); Abrahams (2016); Fevziu et al. (2018)

[6] Only around 6000 Albanians managed to escape to foreign countries between 1944 and 1990. While foreign emigration boomed right at the fall of the regime.

[7] Historical evidence on Italian television watching in Albania are manifold: Dorfles and Gatteschi (1991); Mai (2004); Abrahams (2016); Fevziu et al. (2018) among others. Although in 1973 Italian television watching was forbidden in Albania, people continued to do so regularly. Using World Bank data we compute that around 61% of Albanian household had a TV set in 1990. Data on distribution of TV sets by district in Albania in 1990 can be found in the Online Apppendix.

[8] Interviews of Albanians arriving in Italy in 1990 were conducted, they revealed the extent to which Albanians were familiar with Italian television. More details is available in Appendix A.1.

[9] See Galanxhi et al. (2004). See also Figure A.1 in appendix A.2 which plots yearly emigration flows by destination.

by 2000, an estimated 20% of high-skilled Albanians had left the country (Docquier and Marfouk, 2006; Gërmenji and Milo, 2011; Gëdeshi and King, 2019). Although migration began immediately after the fall of the regime in 1990, its intensity varied with economic and political events: it picked up pace following an economic crisis in 1997, and reached a peak with the war in neighboring Kosovo.

1.5 Data Description

Our analysis builds upon the creation of a novel dataset. We collected information on the Italian TV signal coverage in Albania obtained from RAI along with information about terrain elevation from NASA's Shuttle Radar Topography Mission. For each Albanian municipality, we computed distance measurements and a terrain ruggedness indicator. We then aggregated these datasets at the municipality level, and merged them with the 2005 World Bank Living Standard Measurement Survey for Albania that contains individuals information. Finally, we construct an urban area dataset for Albania for 1986 by classifying NASA satellite images using machine learning techniques.

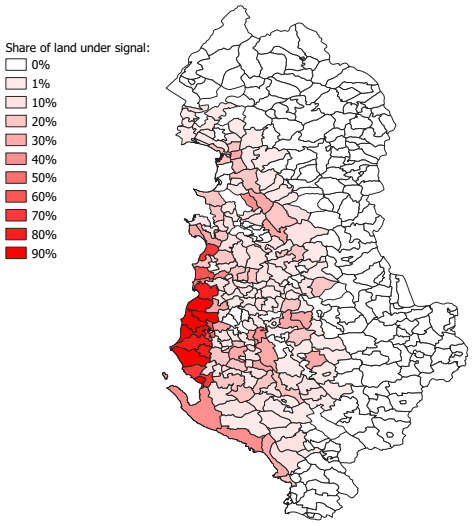
1.5.1 RAI and Geographic Datasets for Albania

We obtained from RAI geographically referenced data on Italian TV signal strength in Albania. The Italian town of Martina Franca is home to the oldest and most powerful Italian TV transmitter able to broadcast all the way into Albania, all other transmitters powerful enough to reach Albania have their signals contained in it. Therefore, we only collected and processed the signal emitted from this antenna. Operational since 1957, the transmitter has not experienced any modifications that altered its power or reception. To compute its signal propagation across the terrain, the RAI uses a standardized forecasting model.[10] We re-classified the dataset of signal quality provided by RAI in two steps. First, to align with RAI's guidelines, we initially transformed signal propagation into a binary dataset, which determines whether Italian TV is accessible for each 100x100 meter grid on the Albanian map. More specifically, we designated Italian television as accessible when the signal quality meets or exceeds a threshold of 55 dBμV/m. Second, we computed for

[10]Prescribed by the International Telecommunication Union, See in particular Recommendation P.526. The model takes into account the diffraction due to the orography of the terrain which reinforces or blocks propagation.

each municipality the share of its area where radio signal is available. Figure 1.1 displays the re-coded TV signal availability across Albanian municipalities.

Figure 1.1: RAI Signal Coverage



Notes: Representation of Albania at the municipality division unit. Signal radio is aggregated at the municipality level to compute the share of area with Italian television access.

We collect topographic characteristics of the terrain from the Shuttle Radar Topography Mission of the NASA which contains information on elevation at a 30x30 meters resolution. From this data we compute the terrain ruggedness index following Riley et al. (1999). We then aggregated both elevation and ruggedness at the municipality level by taking the average over municipality area. We complement this topographic data with distance data, by computing for each municipality the average distance of each of its 30x30 meters cells to Italy, to Greece, to the closest port,[11] and to the antenna in Martina Franca.[12]

1.5.2 2005 Living Standards Measurement Survey Albania

Administered to each household member of 3840 households in 480 primary sample units (geographical census area), the 2005 Living Standards Measurement Sur-

[11] We consider the four most important ports in Albania : Saranda and Vlorë in the south, Durrës in the center, and Shëngjin in the north.

[12] More details is available in the Appendix A.3

vey (LSMS) contains information on 17302 individuals.[13] Restricting the sample to those who were at least 18 years old in 2005, we retain 11040 individuals living across 322 of the 383 Albanian municipalities.[14] As the survey was conducted in 2005, it provides direct information only on non-migrants, however, household heads and spouses are asked to list all their siblings (henceforth, we refer to household heads and spouses that list their siblings as *listing sibling*) and report for each one their demographics, country of living, and year of departure if they migrated. A maximum of seven siblings can be listed, but it's noteworthy that individuals with more than seven siblings only comprise 2% of the total sample. Household members also list their children and spouse living out of the household. We derive two datasets from these sets of questions: (i) one in which each sibling is an observation (27666 obs.); (ii) another in which each child or spouse out of the household is an observation (4714 obs.). Unlike the respondents of the *LSMS*, individuals in these two additional datasets can either reside in Albania or abroad. We thus derive three datasets from the LSMS, the first about the respondents themselves (hereafter, *base dataset*), the second about the household heads and spouses' siblings (hereafter, *siblings dataset*), and the third about household members' children and spouses out of the household (hereafter, *children/spouses dataset*).

Regarding the base dataset, we concentrate our analysis on three types of information: (i) internal migration history since birth; (ii) foreign language proficiency in 1990; (iii) individual education. The exact phrasing of all questions relevant to the analysis of this paper can be found in Appendix A.3.1. Using the internal migration history of respondents, we relocate individuals to their municipality of residence in 1990 and thus to their exposure to Italian television signal before the fall of the regime. Respondents reported their foreign language proficiency in 1990 in Italian, Greek, English or if they had knowledge of "another foreign language". They can answer either 1) Yes, fluently, 2) Yes, some or 3) No. We generate a dummy variable for foreign language proficiency that we code 1 if individuals answer Yes, fluently or Yes, some and 0 if they answer No. Finally, the LSMS records individual's highest education levels, we code a dummy variable equal to 1 if an individual attended university for at least one year. For each individual in the siblings dataset and in the children/spouses dataset, the LSMS includes the country of

[13] Data collection ran between May and early July in 2005. Data and all the material are available at https://microdata.worldbank.org/index.php/catalog/64. Household membership is defined as having been away from the household for less than 6 months during the year preceding the survey.

[14] For underage individuals information is missing.

Table 1.1: 2005 LSMS, Selected Statistics

	Base			Siblings			Children/Spouses		
Variable	All	Men	Women	All	Men	Women	All	Men	Women
Observations	11040	5226	5814	27666	14421	13245	4714	2236	2478
Age Distribution									
25 percentile	24	24	24	37	37	37	27	27	27
50 percentile	40	41	39	45	45	45	33	34	33
75 percentile	53	54	53	55	55	55	40	41	40
Mean	41	41	40	46	46	46	34	34	34
Education									
Primary	54%	49%	58%	.	.	.	52%	53%	51%
Secondary	21%	22%	20%	.	.	.	52%	53%	51%
Vocational	16%	19%	13%	.	.	.	13%	14%	12%
University	9%	10%	9%	.	.	.	9%	8%	10%
Proficiency in 1990									
Italian	5.3%	5.2%	5.3%	.	.	.	7.9%	8.1%	7.7%
Greek	1.9%	2.5%	1.5%	.	.	.	3.1%	4.2%	2.1%
English	4.4%	3.9%	5.0%	.	.	.	3.1%	4.2%	2.1%
Internal Migration									
Before 1990	7.9%	4.7%	11.2%	.	.	.	.	.	.
After 1990	20.6%	16.3%	24.7%	.	.	.	.	.	.
International									
Share migrated	.	.	.	17%	21%	12%	44%	61.3%	28.9%
Before 1990	.	.	.	0.8%	0.6%	1.1%	0.2%	0.2%	0.3 %
Destination									
Italy	.	.	.	32%	32%	33%	39%	41%	36%
Greece	.	.	.	50%	51%	46%	40%	40%	42%
UK	.	.	.	5%	5%	3%	7%	9%	3%
USA	.	.	.	7%	6%	9%	5%	4%	8%
Television									
Ownership rate	62%	.	.	.	.	.	.	.	.

Source: 2005 Living Standard Measurement Survey, World Bank and INSTAT.

residence and the date of emigration. The children/spouses dataset also contains information on foreign language skills in 1990. Where necessary, we attribute the characteristics of their relatives to the individuals in these datasets, in particular their location in 1990 and the highest level of education of the listing sibling to the individuals in the siblings dataset. Note that only children are missing from the siblings dataset and that the children/spouses dataset contains information on a specific sub-population of individuals, namely those who have left the household. Nevertheless, only children account for a minor portion of the population, comprising 6% of the sampled household heads and 3% of the spouses. Furthermore, neither dataset contains individuals from families that have completely emigrated.

Table 1.1 presents descriptive statistics for each dataset. All three datasets are balanced in terms of their sex-ratio, they contain between 49.5% and 53.4% men. With 5.3% of respondents self-declaring their proficiency in Italian, Italian stands as the most widely spoken foreign language in Albania in 1990. English is a close second with 4.4%, and Greek stands third with 1.9% of individuals. In the children/spouses dataset, 28.1% of the sample could speak Italian in 1990, 22.6% could speak Greek, and 15.1% English. Finally, international migration represents 44.2% of the sample of children/spouses out of the household and 16.9% of the sample of siblings. Within the samples, Italy and Greece are the most common destination countries, with 32% of siblings that migrated living in Italy, 50% in Greece; around 39% of children/spouses living abroad are in Italy with an equal share in Greece. In 1990, The LSMS reports that 62% of households were endowed with a TV set.

1.5.3 City-Level Dataset for Albania

We build an urban land cover dataset for Albania for 1986. At a resolution of 30x30 meters, we record for each year and each cell whether it contains urban land or not, and we calculate the proportion of the city that is covered by the Italian TV signal. With this dataset we can calculate the proportion of the urban area of a municipality exposed to the signal in 1986, which we call *Signal II*. This alternative measure has the advantage of estimating exposure only in urban areas, thus avoiding the definition of an exposed municipality when it is mainly exposed in the inhabited area.

1.6 Identification Strategy

A common difficulty in the estimation of a causal effect of signal availability on societal outcomes is the placement of the transmitter. Transmitters are typically placed in strategic locations in order to target specific populations such as densely populated urban areas. In parallel viewers might self-select by relocating to areas where the signal is accessible. This simultaneous selection can substantially bias estimations of causal effects, making the treated population different from the untreated population on unobservables characteristics. The treatment effect on the treated thus differs from the average treatment effect.

The Albanian setting suffers none of these two issues. First the transmitter was placed to satisfy the needs of the Italian population, and no attention was paid to the possibility that the signal might reach Albania, it accidentally did so. Second, emigration was forbidden and internal migration was restricted and centrally managed under the Communist regime, preventing any selection on the Albanian side.[15] Table 1.1 reports that only 0.2% of migrants in the siblings dataset and 0.7% of the migrants in the children/spouses dataset emigrated abroad before 1990. Internal migration tripled from 7.9% of the sample that internally migrated between 1975 and 1990 to 20.6% between 1990 and 2005.

Once controlling for geographic and topographic variables that correlate both with the radio signal exposure and the outcome variables, the exposure to radio signal can be considered as good as random. The controls we consider are: (i) distances to Italy, the transmitter and the nearest port; (ii) topographic data on elevation and ruggedness; (iii) district fixed effects. These controls are potentially correlated with both signal decay and other variables related to our outcomes: migration cost and cultural proximity. Once included, we thus can estimate the effects of residual variations in signal reception due to the topography of the terrain within districts' areas on each outcome variable. We estimate the following specification:

$$y_{i,m,d} = \alpha_0 + \beta \times Sig_m + \gamma \times Dist_m + \theta \times Geo_m + \sum_{d=1}^{36} \alpha_d \times \text{Distr}_d + \varepsilon_{i,m,d} \quad (1.1)$$

Where Sig_m is the share of a municipality's area reached by the TV signal. $Dist_m$ is a vector containing the distances of municipalities to Italy, the nearest port, and the transmitter in Martina Franca. Some specifications also include distance

[15] See Galanxhi et al. (2004) page 9.

to Greece in $Dist_m$. Geo_m controls for the elevation and ruggedness of municipality m. $Distr_d$ are district fixed-effects, such that we measure within a district the differences created between municipalities by the radio signal.[16] $\varepsilon_{i,m,d}$ is the error term. In this specification, β identifies the causal effect of exposure to Italian television on outcome $y_{i,m,d}$. Importantly, it identifies an intent-to-treat effect as we only estimate the effect of exposure to the television signal rather than the one of actually watching Italian television.

We study 2 sets of outcomes: (i) Language proficiency as measured by the self-declared language proficiency in 1990 of individual i living in municipality m of district d in the *base* dataset; (ii) Migration outcome, as measured by whether an individual in the *siblings* dataset lived abroad in 2005 or not. To compute the heterogeneity of the effects we restrict the samples to specific subsets of the population of interest, rather than including a dummy, this approach ensures that fixed-effects and controls are population-specific. Finally, we cluster standard errors at the municipality of residency in 1990 level (i.e. the treatment level) in all regression exercises.

One concern is that municipalities close to the Albanian coastline both concentrate the most TV exposure in the sample and have the lowest distance to either Italy or the transmitter, making it hard to disentangle the effects of distance and TV exposure. If results happen to be sensitive to the exclusion of the municipalities that are the closest to Italy, this might cast doubts on the identification strategy. We address this concern in a number of ways. First, the inclusion of district fixed effects ensures we compare the effect of Italian TV signal between municipalities of the same district, where the distances to Italy are relatively similar. Second, Figure 1.2 in Section 1.9 plots the mean TV signal coverage at different deciles of the distribution of distance to Italy, distance to the closest port and elevation. Although TV signal is concentrated in the first deciles of each distribution, there is considerable variation within and beyond those deciles that allows for meaningful comparisons. Third, in Section 1.9 we go further by showing that results are robust to the exclusion from the sample of the municipalities that are the closest to the ports and the closest to the Greek border. Appendix A.4 proposes balance tests on age and sex ratios using the siblings dataset, confirming that the treated and untreated samples are comparable on observables.

[16] Albania is divided in 36 districts, each district contains 8.6 municipalities on average.

1.7 Results

In the following section we present the results of regressions of the effects of Italian television exposure on Italian language proficiency in the 1990 and on the migration probability of individuals between 1990 and 2005. We show that Italian TV exposure had a sizeable and significant effect on the probability to know Italian in 1990, no effect on the average likelihood to migrate on the Albanian population, but a significant and sizeable effect on the probability to migrate for high skilled individuals, and in particular on the probability to migrate to Italy. Results are paired with placebo tests.

1.7.1 Italian Television and Language Proficiency

This section considers the effect of Italian television on the probability of knowing Italian in 1990. Empirical work highlights the effects of television watching on cognitive outcomes: whether through educational (Kearney and Levine, 2019) or entertainment content (Durante et al., 2019), television has been found to have an influence on human capital accumulation. In our context, we test whether exposure to the Italian language through television pushed Albanians into developing language skills in Italian.

The LSMS is a survey of non-migrants as only individuals who did not migrate until 2005 are eligible to take the survey, hence on this sample we estimate the effect of television access on the acquisition of language skills of individuals that did not emigrate. In line with the literature, if we assume that non-migrants have a lower propensity to learn a foreign language than migrants (Bütikofer and Peri, 2021), then the estimate of the effect of Italian television access on the language proficiency of the non-migrants is a lower bound of the average treatment effect on the full population. We later test this assumption using the children/spouses dataset which contains information on language proficiency and includes migrants and non-migrants, it allows to estimate the effect of Italian television access given migration decisions.

We estimate Equation 1.1 with foreign language proficiency in 1990 as a dummy outcome variable. Table 1.2 columns (1)-(4) present the results of the regression of Italian proficiency on TV signal exposure and three placebo test, using individuals from the base dataset of the LSMS as the sample. In municipalities fully exposed to Italian television, estimated effect in column (1) indicates that the rate of Italian proficiency increased by 7 percentage points, more than double the proportion of

Table 1.2: Italian television effect on foreign language proficiency in 1990

	Base				*Children/Spouse*	
					Abroad	Albania
	Italian (1)	English (2)	Other (3)	Greek (4)	Italian (5)	Italian (6)
Signal	0.070**	0.020	0.006	0.013	0.133**	0.041
	(0.033)	(0.018)	(0.023)	(0.010)	(0.064)	(0.046)
Observations	11040	11040	11040	11040	2088	2626
Clusters	322	322	322	322	233	229
Controls:						
Common: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness						
Greek Community	N	N	N	Y	N	N

Notes: The table reports OLS estimates of the effect of exposure to Italian TV on foreign language proficiency in 1990. (1)-(4) use the base dataset, (5)-(6) the children/spouse dataset, in specification (5) the ones living abroad and in (6) the ones living in Albania. The dependent variable is the reported capability of speaking Italian, English, Other (category any other language), and Greek in 1990 coded as a dummy. The main explanatory variable, Signal, is the share of a municipality's area with access to Italian TV. Controls for Greek community in specification (4) include distance to Greece and dummies for: (i) Greek ethnicity, (ii) orthodox religion; (iii) Greek as maternal language; (iv) speaks Greek daily at home; (v) speaks Greek in the community. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Italian speakers in Albania. The estimate is significant at the 5% level and economically sizeable. As explained above, since the sample only include non-migrants, this estimate is a lower bound of the average treatment effect. Additionally, as we can only measure television exposure, not television watching, we can only estimate an intention-to-treat effect, lower than the average treatment effect.

Columns (2) to (4) of Table 1.2 report the results of placebo tests, we check that Italian television exposure did not cause an increase in proficiency of other languages. As expected, the coefficients are small and insignificant. In the case of Greek language proficiency, we added controls related to whether individuals in the sample belong to the Greek diaspora present in Albania, additional controls

include dummies for Greek ethnicity, Orthodox religion, and Greek spoken daily at home. We also included a distance variable that measures for each municipality its distance to the Greek border.[17]

To overcome the limitation of the base LSMS sample, which only contains non-migrants, we exploit the dataset of children/spouses which contains household members' spouses and children that no longer live in the households, and can either be international migrants or still in Albania. Regressions (5) and (6) present the estimates of regressions on the sub-sample of children/spouses that respectively live abroad and in Albania. We find a sizeable effect of exposure on Italian proficiency on the sub-sample that lives abroad of 13 additional percentage points in fully exposed municipalities, triple the estimate on the sample of non-migrants. This result approximates the intent-to-treat effect of Italian television exposure on migrants, it confirms that migrants have a higher propensity to learn a foreign language than non-migrants (Bütikofer and Peri, 2021).

The lower bound of 7 percentage points increase we estimate indicates that the impact of television access on rates of Italian proficiency in Albania was considerable, more than doubling the rate of Italian proficiency. We thus find that exposure to foreign media can be used as an effective tool to foster foreign language proficiency. Given empirical results that link linguistic proximity and emigration (Belot and Ederveen, 2012; Adsera and Ferrer, 2015), we expect television access to have impacted patterns of emigration to Italy through its impact on language proficiency.

1.7.2 Italian Television and Brain Drain

In this section, we investigate the effect of Italian television on the emigration decisions of Albanians between 1990 to 2005. The literature underlined the penalty that linguistic differences represent for highly-skilled migrants (Adsera and Ferrer, 2015) owing to the complementarity between language and skill (Chiswick, 1995; Berman et al., 2003). The seminal paper of Borjas (1987) also underlined the importance of such mechanisms in driving the self-selection of migrants. As a consequence, we expect the effect of Italian television to have differed across skill groups.

To conduct this investigation, we resort to the siblings dataset. As discussed in Section 1.5, the LSMS respondents are all non-migrants, we thus exploited a sample composed of their siblings, that can either be migrants or non-migrants.

[17] The case of Greek, owing to the particular history between the two countries, is further discussed in Appendix A.5.

We assume that siblings of respondents were living in the same municipality as respondents in 1990, consistent with the low internal migration rates characterizing Albania before 1990. As international migration was forbidden prior to 1990 (see Section 1.4) individuals residing in a foreign country in 2005 migrated between 1990 and 2005. We attribute to siblings the human capital of their listing siblings: we assume that education levels of siblings were highly correlated. Specifically, we define a sibling as high skilled if her listing sibling attended university for at least one year. We exploit the identification strategy described in Section 1.6. Standard errors are clustered at the municipality of residency in 1990 level (treatment level).

Table 1.3: Effect of Italian Television Exposure on Probability to Migrate

	Siblings dataset			
	Abroad (1)	Abroad (2)	Italy (3)	Greece (4)
Signal	-0.002	0.244***	0.131**	0.0518
	(0.031)	(0.074)	(0.059)	(0.074)
Sample	Full	High Skill	High Skill	High Skill
Observations	27666	2153	2153	2153
Clusters	310	128	128	128
Common Controls: District F.E., Distance to Italy, to transmitter, to Port, Elevation, Ruggedness				

Notes: The table reports OLS estimates of the effect of exposure to Italian television on the probability to reside abroad. The outcome variable is a dummy taking value 1 if abroad for columns (1) and (2), 1 if in Italy for column (3) and 1 if in Greece for column (4), and 0 otherwise. All specifications use the siblings dataset (see Data Section 1.5). Specifications (2)-(4) restrict the sample to siblings of individuals that attended university for at least one year. Signal is the share of a municipality's area exposed to Italian television signal. Standard errors are clustered at the municipality of residency in 1990 level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 1.3 column (1) reports the estimate for the effect of TV signal exposure on individuals' probability to migrate abroad, we do not find evidence of an effect. In column (2) we subset for the population of individuals whose listing sibling attended university for at least one year and repeat the estimation of Equation 1.1.

We find an economically and statistically significant effect: Italian television signal increased the migration probabilities of fully exposed individuals by 24 percentage points. We stress that the sample of high-skill individuals represents only 10% of the total sample in specification (1), implying that the effect is attenuated in the full sample. The positive effect we estimate on high-skill individuals is not paralleled by a significant negative effect on individuals with other education levels.[18] Finally, we test whether the destination of the emigration of the high-skilled is Italy, columns (3) and (4) test for emigration by destination. Although reduced in magnitude, Italian television signal access significantly increased emigration towards Italy, and left emigration rates towards Greece unchanged. In Section 1.9 we successfully test our results with alternative specifications of TV signal and human capital measures.

To move from this reduced-form estimate to the effect of language proficiency on emigration probability, we would need to perform an instrumental variable regression. However, we do not have information on the Italian language proficiency for siblings, and would thus need to rely for the first-stage on the results of the regressions of foreign language proficiency on the sample of non-migrants. It would underestimate the effect of TV on language proficiency, inducing an overestimation of effects in the second stage. Nonetheless, since the effect of TV exposure on Italian proficiency is necessarily bounded between 0 and 1, we do know that the reduced form estimate is necessarily a lower bound of the effect of language proficiency on the migration probabilities of the highly skilled.[19] Given that the reduced-form estimate already shows a substantial 20 percentage point increase in the likelihood of emigration, we can confidently assert that foreign language proficiency strongly enhances the migration probability of highly skilled individuals.

Taken together, our results imply that Italian television accentuated the brain drain towards Italy. Television access pushed many educated people into emigration towards Italy, thus increasing the positive selection of emigrants and contributing to the brain drain. Previous research investigating the impact of the media on migration behaviour emphasized the role of the media as a source of information (Farré and Fasani, 2013; Pesando et al., 2021; Adema et al., 2022). In this setting, given results on language proficiency, we expect the language-skill complementarity characterizing highly-skilled individuals to have played an important role in

[18] Results on the rest of the sample are available in Appendix A.6.3.

[19] In appendix A.6, we nonetheless test for the two sample instrumental variable regression. Results are too imprecise to provide meaningful information.

raising their returns to emigration (Chiswick, 1995; Berman et al., 2003). We posit that language proficiency is the main mechanism through which Italian television exposure increased the emigration of the educated. In the next section, we discuss the exclusion restriction to our identification strategy.

1.8 Exclusion Restrictions

1.8.1 Italian Television, Competing Channels to Language Proficiency

The most widely discussed channel in the literature is the one of information: migrants' expectations about income abroad can be biased (McKenzie et al., 2013), television and media might correct these expectations by providing valuable information about life abroad (Farré and Fasani, 2013; Adema et al., 2022). Applied to the Albanian context, Italian television would have provided high-skill individuals with information about economic opportunities in Italy. In this section, we provide evidence to rule out the role of this competing channel.

Historical sources emphasize that Albanians were watching entertainment programs on Italian television, and data confirms that picture. Dorfles and Gatteschi (1991) reports results of interviews conducted in March 1991 on 311 Italian speaking Albanian migrants just arrived in Italy. Of the people interviewed, 301 declared they were watching Italian television in Albania, they were further asked which Italian television programs they would usually watch. The overwhelming majority of programs listed, 93%, are entertainment programs, only 7% of listed programs were news shows. In Farré and Fasani (2013), it is precisely news content which induced potential migrants to revise their beliefs. As interviews reveal, Albanians mainly watched entertainment programs: they were not being provided with useful information thanks to Italian television.[20]

We dispel further concerns about the contribution of the informational channel by exploiting the migration questionnaire of the LSMS. Members of surveyed households are all asked whether they migrated for at least one month since the age of 16 (since respondents are all in Albania, they would by definition be temporary migration episodes). Those who responded positively were subsequently asked "*who provided information on where to go and/or how to find work during this first migration episode*". Respondents can choose their answer from a list including the item *TV, radio, newspaper or book*. Table 1.4 presents the distribution

[20] A detailed presentation of the results of these interviews in available in Appendix A.1.

of answers: only 1% of individuals chose this item. Even though the sample interviewed is one of return migrants, it is informative of what migrants themselves would have answered, and indicates further that television was not used as a source of information.

Beside information, watching entertainment television could have led Albanians to form an idealized view of life in Italy as suggested in Mai (2004). We would expect such an effect to have been homogeneous across skill groups, it could nonetheless turn heterogeneous in our sample if low-skilled individuals faced liquidity constraint preventing them from financing migration project. Alternatively, it could also be that TV ownership was correlated across skill-groups. Table 1.4 addresses these concerns. First, it shows that individuals migrated across all education groups with only small differences in emigration rates, indicating that individuals with lower education levels were not necessarily liquidity constrained. The same is true for television ownership: although its rate increases with education, TV sets were widespread across education groups.[21] This evidence suggests that how Italian television shaped beliefs did not impact migration patterns.

1.8.2 Italian Television and Returns to Education

Another competing mechanism is that television watching raised the return to education. Shrestha (2017) shows that the possibility to migrate can in some context raise returns to education, thus increasing the average education of the population. If this is the case in the Albanian context, university educated individuals in municipalities with Italian TV access might differ on unobservables from university educated individuals who lived in municipalities without such access because they were pushed into accumulating more human capital by TV access, and these unobservables may drive our results. To remove these unobservables from the sample, we restrict it to individuals that completed their education prior to the fall of the regime in 1990. The only way in which Italian TV might have increased education returns is by raising the returns on emigration, as migration was forbidden before 1990, this effect must have been absent. We therefore estimate the siblings' hypothetical age of graduation, defined as the age each sibling would have graduated if they had completed their education at the same age as the listed sibling. We then filter out from the sample individuals with hypothetical year of graduation posterior to 1990. Specification (5) of Table 1.6 confirms our baseline results on this

[21] In addition, historical sources report that group viewing of Italian television were regular and frequent, people did not need to own a TV to watch Italian television regularly.

Table 1.4: 2005 LSMS, Selected Statistics

Variable	Base Dataset Share
Information provider:	
Family/relatives in Albania	0.03
Family/relatives abroad	0.30
Friends in Albania	0.14
Friends abroad	0.41
Previous personal experience	0.08
Neighbours	0.02
TV, radio, newspapers	0.01
Internet	0
Others	0.01
Owns a TV in 1990 by education:	
Primary or less	0.57
Secondary	0.68
Vocational	0.68
University	0.78
Emigrated by education:	
Primary or less	0.14
Secondary	0.20
Vocational	0.20
University	0.21

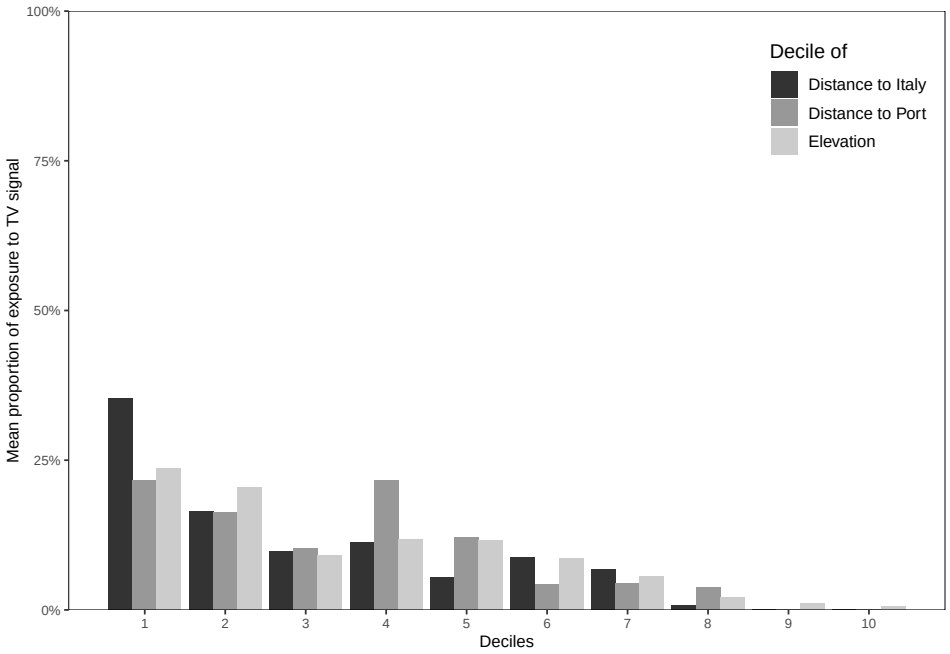
subsample, hence, even among individuals that accumulated human capital before 1990, when Italian television could not have raised the returns to education, we find the same effect on the emigration of the educated.

The results presented thus far suggest that neither the informational channel nor the belief channel played any role in fostering emigration towards Italy of the highly-skilled Albanians. We conclude this section by underlining the role played by language proficiency which, given the language-skill complementarity (Chiswick, 1995; Berman et al., 2003), raised the returns to migration of high skilled individuals.

1.9 Robustness

This section presents robustness tests of our results. First, we show that our results do not depend on the higher exposure of the Albanian coastal areas to the

Figure 1.2: Radio Signal and topographic data



Notes: Each bar represents the mean share of municipalities' areas under the signal for municipalities that can be found in the decile of the relevant topographic variable considered.

Italian television signal. Second, we show that our results are not sensitive to the exclusion from the sample municipalities closest to ports and closest to Greece, where migration costs were low. Finally we test that results are robust to alternative identifications of high-skilled individuals, and to different definitions of signal exposures.

Since television access is concentrated in the coastal areas, a concern surrounding the identification strategy is the high correlation between signal power and distance measurements. As the latter are directly related to migration costs (the further away from Italy, the more complicated to migrate), high levels of correlations might result in spurious estimations. Figure 1.2 comes to alleviate this concern: although most of TV exposure is concentrated in municipalities within the first deciles of the distances to the closest port and to Italy, there is significant variation in signal exposure between municipalities in all deciles up to the 7^{th}.

Another concern is that the inclusion of municipalities of coastal areas and bordering Greece puts in the sample individuals unlikely to be impacted by Ital-

ian television: their migration costs might be so low that there is little role left for television. In Table 1.5, we limit the sample to highly-skilled individuals that lived more than 30km from a port (1st quartile of distance to ports) and more than 48km for the Greek border (1st decile of distance to Greece). Results dispel all doubts related to spurious estimation: the coefficients of interests are still precisely estimated, significant at the 5% level for migration abroad, and at the 1% level for migration to Italy. It is worth noting that in this regression exercise, the effect of the signal on migration and the effect on migration to Italy collapse to the same point estimate. It suggests that the difference between the two coefficients we observed in Table 1.3 is due to the presence of *always-takers* who would have migrated even if they had not been exposed to the signal.

Table 1.5: Effect of Italian Television Exposure on Migration Decision. Sensitivity to Coastal Areas and the Greek Border.

	Siblings dataset		
	Abroad	Italy	Greece
	(1)	(2)	(3)
Signal	0.168**	0.168***	-0.0334
	(0.0675)	(0.0519)	(0.0792)
Sample	High Skill	High Skill	High Skill
Observations	1476	1476	1476
Clusters	72	72	72
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness			

Notes: The table reports OLS estimates of the effect of exposure to Italian Television on the probability to be abroad. Outcome variable is a dummy taking value 1 if abroad (1), in Italy (2), in Greece (3). All specifications exploit the siblings dataset (Section 1.5), restrict the sample to individuals whose *listing sibling* attended university for at least one year. Signal is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Municipalities that are less than 30 Km of distance (1st quartile of distance to ports) and less than 48 km from the Greek border are removed from the sample. Clustered standard errors at the municipality level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

As an additional robustness check, we test whether our results depends on the

Table 1.6: Effect of Italian Television Exposure on Migration Probabilities: Alternative Definitions.

	Siblings Dataset						
	Abroad (1)	Italy (2)	Abroad (3)	Italy (4)	Abroad (5)	Italy (6)	Abroad (7)
Signal II	0.148**	0.139**					
	(0.0742)	(0.0638)					
Signal			0.154**	0.0819*	0.177***	0.104**	0.221**
			(0.0690)	(0.0472)	(0.0592)	(0.0450)	(0.0958)
Sample	H. Skill	H. Skill	Small Fam.	Small Fam.	Wealthy	Wealthy	H. Skill $\leq$ 1990
Observations	2153	2153	2449	2449	4043	4043	1510
Clusters	128	128	243	243	240	240	107
Common Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness							

Notes: The table reports OLS estimates of the effect of exposure to Italian TV on the probability to be abroad and in Italy. It replicates specifications and results of Table 1.3 with alternative definitions of signal exposure and human capital using the siblings dataset. In particular, specifications (1) and (2) repeat specifications (2) and (3) of Table 1.3, but use 1986 municipalities' share of urban area exposed to the signal as explanatory variable instead of the usual signal definition. Specifications (3)-(6) exploit the usual signal variable but change the definitions of high skilled individuals. (3)-(4) subset the sample by for family with less than 4 children, while (5) and (6) subsets for individuals that lived in an apartment in thev 4th quartile of the distribution of the number of rooms per person in 1990. Specification (7) identifies an individuals as high skilled individuals if she attended university and completed her education prior to 1990. Clustered standard errors at the municipality of residency in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

specifications of either TV signal exposure or the measures of human capital. We first vary the measure of TV signal exposure, while we previously used the share of a municipality's land exposed to the signal, we here use the share of the municipality's urban area (in 1986) exposed to the signal. This avoids accounting for signal reaching inhabited rural areas. Results of this exercise are presented in table 1.6 columns (1) and (2). With this refined measure, using as dependent variable migration abroad or specifically to Italy yield the same results of 14 percentage points, thereby confirming earlier results.

In the main specification, we identify an individual as highly skilled if her listing sibling attended university at least one year. In this robustness check, we use alternatively family and dwelling dimension, characteristics we can measure for the sibling herself, as a proxy for education. Large families are more likely to come from agrarian backgrounds, where it is less likely that children can be sent to university. We thus split the sample according to family size, identifying a family as small if it is composed of less than four children. Columns (3) and (4) implement this sample split, repeating main specification (1.1). Results are

qualitatively similar, although less precise. In Appendix A.6.3 we show that the effect of TV exposure decreases in the family dimension. We define the dwelling dimension as the ratio of the number of rooms to the family dimension, yielding the number of rooms per person. Much like for family dimension, we assume that housing size is correlated with education. In (5) and (6), we subset the siblings sample for the fourth quartile of the distribution of housing dimension and confirm our baseline results. Appendix A.6.3 shows the alternative regressions for smaller dwellings. These robustness exercises confirm that our results do not hinge on the definition of either TV signal access or on human capital.

1.10 Conclusion

How much does foreign language proficiency affect individuals' migration probabilities? Answering this question is relevant for policy makers deciding whether to promote plurilingualism in society. It is also relevant for understanding the causes of brain drain. So far, this question has only been addressed by observational studies, unable to address the inherent self-selection issues that characterize these settings.

In this paper, we exploit a natural experiment that occurred in Albania in the second half of the twentieth century to assess the causal effect of foreign language proficiency on high-skilled migration. We show that as good as random exposure to Italian television increased by at least 7 percentage points Italian language proficiency and by 24 percentage points the likelihood of migration of high-skilled individuals in fully exposed municipalities. We interpret the effect of signal exposure on foreign migration as the effect of higher language proficiency and rule out competing channels.

While our study contributes novel insights by establishing the causal impact of foreign language proficiency on the migration of highly-educated individuals and documenting the effects of foreign media on language proficiency, there are limitations to our analysis. It is restricted to estimating a reduced-form equation of the impact of language proficiency on emigration, as we cannot estimate an instrumental variable regression due to data limitations. Therefore, we provide a lower bound estimate of the impact of TV exposure on language proficiency, and as a result, we can only offer a lower bound estimate of the effect of language proficiency on the likelihood of emigration.

The economic literature still presents diverging results as to the effects of brain

drain on the economy Shrestha (2017); Anelli et al. (2023), we leave for further research the evaluation of the impact of the Albanian brain drain on Albania's economic development.

Chapter 2

Is There a Devaluation of Degrees? Unobserved Heterogeneity in Returns to Education and Early Experience

Co-written with **Robert Gary-Bobo** and **Marion Goussé**

2.1 Abstract

We show that the expected real wages commanded by some higher-education degrees decreased in absolute terms in France, in the past two decades, and that this drop is not due to adverse selection. To study the returns to degrees and experience, we assume the existence of a finite number of latent types and estimate a finite-mixture model. Each type has its own log-wage equation, experience-accumulation and education-choice equation. This allows us to decompose the treatment effects of education as an average of type-dependent effects. We then show that some unobserved types experienced a real-wage drop while others benefited from an increase, with the same degree. The observed "flattening" of returns to experience is also heterogeneous. In the case of Master degrees, the estimated distribution of latent types indicates that student selection improved with time, in spite of the fact that the number of graduates increased substantially. An excess supply of graduates might therefore be a likely explanation for the devaluation of Master's degrees.

2.2 Introduction

In the past decades, the enrollment of Universities and Colleges has grown substantially in many countries.[1] The growth in the number of persons who reached tertiary education between 2010 and 2020 is impressive.[2] Various authors have argued that this growth has the potential to cause an excess supply of graduates, and hence a decrease in the real wages of College graduates relative to high-school graduates (*i.e.*, a drop in *college premia*).[3]

In the following, we study the returns to education and experience of young French men. We first show that some higher-education degrees, mainly the Master degrees, command on average a smaller real wage in the period 2010-2017, as compared to the period 1998-2005, and we propose a study of the causes of this drop. We model the unobserved heterogeneity with a system of latent types in order to address the important endogeneity problems of education and experience. We find that the observed drop in the return to Master's degrees is not due to a deterioration of the average quality of students (*i.e.*, adverse selection), in spite of enrollment growth.

For convenience, we define the *devaluation* of a given degree as an absolute decrease in the average real-wage of the holders of this degree (all other things being equal). Devaluation is commonly measured in relative terms, taking the form of a drop in the College wage premium, which we also observe, but for some categories of degrees, the drop can be absolute, as we will see below. When a devaluation of degrees is observed, it is an open question to disentangle the possible effects of an excess supply of graduates, the change due to a lesser quality of teaching, and finally, the variation caused by a less favorable selection of students. In the present article, using an econometric model with unobserved individual types, allowing for unobserved student heterogeneity, we conclude that, in the past 25 years, in France, the observed devaluation of Master's degrees does not seem to be due to a deterioration of the selection of students. On the contrary, it seems that the quality of selected students has improved, while at the same time, the number of graduates did increase substantially. We also show that the decrease in average real wages, conditional on some degree or level of education, is an average of heterogeneous variations through time: some unobserved types of students do not suffer from de-

[1] A "big push" occurred. According to OECD figures, in the United States, 7.7% of the population aged 25 or more had graduated from College in 1960, as compared to 37.5% in 2020.

[2] *Education at a Glance*, oecd-ilibrary.org, 2022.

[3] See our discussion of the literature below.

valuation, while the real wage of others decreases. These variations are in turn due to changes in returns to education and returns to experience that are themselves heterogeneous.

In essence, the present article builds and estimates a model explaining individual wages, individual employment rates and education choices simultaneously with the help of panel data. We assume the existence of a finite number of latent individual types. Each type has a specific (*i.e.*, type-dependent) log-wage equation, a specific employment-rate equation and a specific discrete-choice model describing educational investment. In other words, the model is the product of three finite-mixture models for respectively, wages, employment and education, describing the accumulation of effective experience and the returns to experience of each latent type, as well as type-dependent returns to degrees.

We then assume that error terms are normally distributed in the log-wage and employment equations (resp., extreme-value distributed in the education-choice equation) *conditional* on observable and unobservable characteristics. The model is flexible: there are no cross-equation or cross-type restrictions and it is well-known that any smooth distribution of wages can be approximated, to any desired degree of precision, by a mixture of normal distributions. Thus, we assume that the endogeneity problems, typically arising in standard econometric regressions such as the Mincer equation, are entirely driven by the unobserved types: error terms are assumed independent of controls and types. In a nutshell, in this type of structure, identification of type-dependent parameters and of the distribution of types essentially relies on the panel structure, since each individual is typically observed more than twice, but we do not use instruments for identification.[4] We discuss the conditions for *nonparametric* identification of the model below, relying on results giving conditions for the nonparametric identification of finite latent structures.

The model is estimated by straightforward likelihood maximization. Good preliminary estimates are generated by means of a sequential EM algorithm. Given that almost all model coefficients are type-dependent, the number of parameters increases quickly with the number of latent types, but we estimated a model with 3 types by Maximum Likelihood. We discuss the choice of the number of types and show that three types is a reasonable choice. In particular, the three types provide a surprisingly good *classification* of individuals (with low entropy). The gains of an additional type are small.

[4]Instruments, could be added without difficulty, but we would then allow for a type-dependent impact of the instruments.

An important output of finite mixture models is the probability of belonging to a given type, conditional on the individual's observed characteristics (hereafter the *individual posterior probabilities of types*). These probabilities show the most likely type of each individual. Heckman and his co-authors (see, *e.g.*, Heckman and Vytlacil (2005), and more recently, Heckman et al. (2018)), propose to derive and compute *policy-relevant parameters* from the output of a structural model, in the presence of treatment-response heterogeneity.[5] Our model's estimated coefficients and the posterior probabilities of types allow us to compute a number of policy-relevant parameters very easily. Posterior probabilities can be used as a system of weights to evaluate treatment effects. In particular, considering education as a treatment, when wages are the outcome, we can compute the ATT and ATE[6] of a certain level of education (or of a certain category of degree) and compare the two. The estimated model also allows us to compute ATEs conditional on the latent type and to uncover the heterogeneity of effects across types.

We estimate the model on a rich panel of young French workers: the *Generation surveys*. In these data, we follow the first seven years of career of three cohorts of workers. Each cohort is defined by the year during which the worker left the educational system, namely, 1998, 2004 and 2010. We show the existence of heterogeneity in the returns to experience across types. We find an erosion of returns to effective experience on average (and a corresponding flattening of wage curves), but we also find that the returns to experience of one of the types increased while that of the other types decreased.[7] We believe that the observed absolute devaluation of University Master's degrees, in France, is most likely due to an excess supply of graduates, because we find that the selection of students has improved with time. In contrast, we find that in the French business schools, the enrollment of which has also grown substantially, the quality of student selection has decreased with time. Then, we use simulations of the model to generate fictitious careers and compute discounted expected earnings over the first 7 years of career, type by type. Simulation results typically confirm the findings and give a synthetic view of degree devaluation.

Our model is relatively simple and easy to estimate. It can be called semi-

[5] We do not use the same model as Heckman and his co-authors, but our philosophy is similar, and the influence is direct.

[6] *Average treatment effect on the treated* and *Average treatment effect*, resp.

[7] It follows that the devaluation of degrees may be underestimated for at least two reasons: firstly because it is a dynamic phenomenon, taking a few years of career to reveal itself fully, and secondly, because a subset of types do suffer from devaluation while others do not.

structural: we do not explicitly model the sequential choice process of individuals (Heckman et al. (2018) describe their work as developing "a methodological middle ground between the reduced-form treatment approach and the fully structural dynamic discrete choice approach"). Our description of education choices is essentially static and our employment equation (experience-accumulation model) is a kind of reduced form.[8] The types that we find are easily interpretable: there is an obvious ranking of types in terms of returns to education.

Literature. Empirical work confirms that returns to higher education have recently decreased. See, for instance, Valletta (2018), Emmons et al. (2019). In the United States, the 80s and 90s have been characterized by the rise of the College skill premium. Increased inequalities have been attributed to skill-biased technical change. The work of Katz and Murphy (*i.e.*, Katz and Murphy (1992)) shows that a "standard" model is able to capture the evolution of the hierarchy of wages as the result of an increased demand of employers for graduates (and for the employment of women). Fluctuations of the skill premium are directly related to the supply of graduates. Card and Lemieux (2001) have then showed differences in the evolution of skill premia across age groups and emphasized that the main force favoring the relative wages of younger graduates is the smaller growth of their number in the generations born after 1950 in the US, the UK and Canada. Goldin and Katz (2008) propose a historical view of wages over more than a century in the US.[9] This line of research has led to an analysis of labor-market *polarization*, and the so-called "Ricardian" model of the allocation of skills to tasks, allowing a study of occupational downgrading (see, *e.g.*, Acemoglu (1999), Autor et al. (2008), Acemoglu and Autor (2011)).

In the UK, Blundell et al. (2022) propose an explanation for the fact that the proportion of UK workers with university degrees tripled between 1993 and 2015 while simultaneously the time trend in the college wage premium remained flat: during the period, firms opted for more decentralized organization forms, UK firms took advantage of an increased supply of graduates and chose to pick up the technologies and organizational forms already developed in the US. Ichino et al. (2022) study the higher education expansion in the UK from 1960 to 2004 with the help of a general equilibrium Roy model. They find that the expansion is associated with a decline of the average intelligence of graduates and that it mainly benefited

[8] But of course, simplicity comes with some benefits in terms of tractability and interpretation.

[9] According to Autor et al. (2020): "The largest part of increased wage variance in the twenty-first century comes from rising inequality among college graduates...".

relatively less intelligent students from advantaged socioeconomic backgrounds.

Until the turn of the millenium, facts were giving the impression that the evolution of wages and skill-premia had been different in France and in the US, with no development of inequalities due to higher education in the former country. Indeed, the work of Verdugo (2014) shows that France has experienced a *great compression* of the hierarchy of wages until 2008. But, finally, it may be that similar phenomena have been at work in the two countries and in the recent years. In the United States, Beaudry et al. (2014, 2016) have shown the existence of a trend shift around the year 2000. The share of the working population commonly allocated to cognitive-task occupations has ceased to grow at the turn of the century, while the share of graduates was still increasing. The result was an increased probability of occupational downgrading, with various adverse consequences for the less qualified workers. After 2000, the wage curves of the 4-year College graduates have "flattened" and the starting wages went down, and these facts cannot simply be explained by the business cycle. The situation of France is similar.

Studying composition effects, Carneiro and Lee (2011) show that enrollment growth is likely to have caused a decrease in the quality of student selection, explaining a drop of 6% in the College skill-premium between 1960 and 2000, in the United States.[10] Belzil and Hansen (2020) reach similar conclusions, comparing the 1979 and 1997 cohorts of the NLSY survey, using structural econometric methods. Ashworth et al. (2021) use the same NLSY data, and study closely related questions with the help of a structural model with a latent factor structure.

Returns to experience have recently been the object of renewed interest: see Dustmann and Meghir (2005), Kambourov and Manovskii (2009) and Jeong et al. (2015). On the dynamics of wages over the life-cycle, see *e.g.*, Huggett et al. (2011), Magnac et al. (2018), Guvenen et al. (2021). Our analysis shows that the age-earnings profiles of individuals with different latent types (and with different levels of human capital) also have different slopes as in — for instance — Guvenen (2007).

For a general treatment of finite mixture models, see McLachlan and Peel (2000), Bouveyron et al. (2019). The estimation methods used here have been employed in various contributions. Discrete or discretized latent structures are not a novelty in economics, and go back (at least) to Heckman and Singer (1984).

[10] Their result relies on an identification assumption: there are College-enrollment differences in the individuals' regions of birth that can be exploited to disentangle the effect of quantity from that of quality. See also Carneiro et al. (2011), who estimate the *marginal treatment effect* of College.

The sequential EM algorithm that we use to obtain preliminary estimates has been proposed by Arcidiacono and Jones (2003) and applied by several researchers[11]

2.3 Context and Data

In this section, we first briefly describe the French education system and present the data.

2.3.1 The French context

In France, (as in many other countries) the share of higher-education graduates has constantly grown in the past decades. In 2012, the share of higher-education graduates, including the French equivalent of the associate's and two-year vocational degrees, reaches 42% in the 25-29 age bracket, while this share is only 12.5% in the 60-64 age group. The number of higher-education students has reached 2.73 million in 2019-2020.[12] Between 1990 and 2015, the overall rate of growth of enrollment in higher-education institutions reaches 37%. The growth in enrollment has the potential to flood the labor market with graduates and there are concerns that an excess supply of Masters would cause a drop in their wages. In 2002, French universities have implemented the BMD reform (*i.e.*, the *Bachelor-master-doctorate* reform), *i.e.*, a set of measures adapting the French higher education system to European standards. The reform has set up an architecture based on three academic levels: Bachelor (*i.e., Licence*), Master, PhD (*i.e., Doctorat*).[13] The system is described by Fig. 2.1.

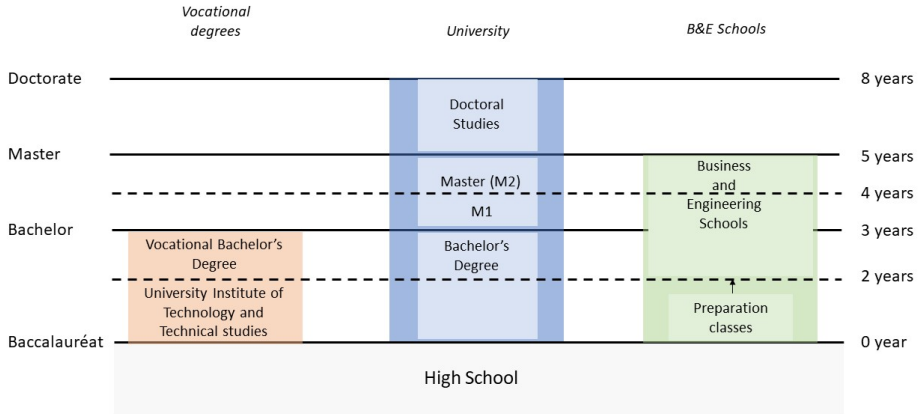
After high-school graduation (*i.e., baccalauréat*), typically at 18, students may go to work or continue studying. This depends to a large extent on the type of baccalauréat, that can be vocational or general. There is a group of vocational degrees requiring two or three years of education that can be compared to Associate's degrees in America. Undergraduate studies in universities lead to a Bachelor's de-

[11] See, in particular, Beffy et al. (2012), Arcidiacono et al. (2016). In addition, Gary-Bobo et al. (2016), Cassagneau-Francis et al. (2021), and Cassagneau-Francis (2021) present other applications of finite mixtures. Finally, see the recent manuscript of Corblet (2022), exploring the same data sets, but with other methods.

[12] Figures published online by the French Ministry of Higher Education and Research, *i.e., Ministère de l'enseignement supérieur de la recherche et de l'innovation*.

[13] Higher education in France is now structured by European standards. Before this reform, the French system was not very different, and it is easy to find a correspondence between the pre-reform and post-reform degrees. The division of institutions in three categories: universities, vocational institutes, and schools has survived.

Figure 2.1: **The French Higher-Education System**



gree after three years of College. The students who continue after three years in universities typically enter a two-year Master program. We distinguish the first year (called M1, standing for Master 1), 4 years after high-school graduation, from the second year (called M2, standing for Master 2) and requiring 5 years of study. The reason for this distinction is selective admissions. Until recently, the French public universities were not allowed to select students at the entry of M1 years. The tradition was that selective admission was permitted only at the entry of the second, M2 year (and this was the rule applied during our observation period). In the period covered by our data, the M1 is still not selective in principle. Yet, some universities used capacity constraints to limit admissions. But in this system, M2 Master graduates are obviously special. Finally, there also exists engineering schools and business schools that are typically independent institutions and have nothing to do with universities. Some are public, some are private (mainly non-profit) institutions. The best such schools deliver a degree after five years of study, but the first two years are devoted to preparation classes. Admission is typically selective, sometimes very selective, in all French higher-education schools. They admit students after a competitive entry exam. The schools' degrees are equivalent to Master's M2 degrees but the selection of students is of course much more rigorous in schools than in universities, at least in principle.[14] This is the reason why we single out the business and engineering school categories, including only students with at least 5 years of study after high-school graduation in this category.

[14] Some engineering or business schools — not the best — admit students directly after high-school graduation. In addition, business schools recently developed 3-year bachelor programs that are less selective, enrolling students after high-school.

Aggregation of degrees In the following, we aggregate the highest degrees of individuals in 5 categories: 1°) *Below High-School Degree*, including dropouts without any certificate and secondary vocational certificates[15]; 2°) *High-School Degrees*, including all students for whom the *baccalauréat* is the highest degree. Many of these individuals in fact earned a vocational certificate, the *baccalauréat professionnel*, and in contrast, many of the classical baccalaureates have been enrolled in various higher-education institutions and therefore eventually earned a more advanced degree; 3°) *Some College and Bachelors* includes all the students whose highest achievement is the equivalent of an Associate degree[16], plus all the bachelors, *i.e.*, the French *Licence* and the M1, *i.e.*, the first year of master programs; 4°) Master degrees, typically the degree of a two-year graduate program requiring 5 years of study (M2); 5°) The degrees of all business and engineering schools, also requiring 5 years.

2.3.2 Data; CEREQ Generation Surveys

For the estimation work presented here, we stacked three large samples of young male workers. The samples, called Generation Surveys (*i.e.*, *Enquêtes Génération*) are produced by a French public institution called CEREQ.[17] Since 1992, every 5 years, the CEREQ draws a large representative sample of individuals who all left the educational system during the same year, with a large variety of educational achievement levels.[18] We will consider the CEREQ Generation surveys of 1998, 2004 and 2010 only.[19] The workers are followed during 7 years; the labor market experience of each worker is tracked, month by month, by means of interviews; questions are asked after 3 years, 5 years and 7 years. The data takes the following form : a listing of individuals with, for each, a long list of possible control variables (including family-background characteristics) and a fine description of degrees and certificates. For each individual, we also have a list of employment and unemployment spells, giving the monthly wage at the beginning and the end

[15] In particular, the CAP, *i.e., certificat d'aptitude professionnelle*, the BEP, *i.e., brevet d'études professionnelles*.

[16] The degrees of the IUT, *i.e., Instituts Universitaires de Technologie*, called DUT, of the STS, *i.e., Sections de techniciens supérieurs*, called BTS, and other vocational degrees requiring less than 3 years of study.

[17] The CEREQ (*i.e., Centre d'études et de recherches sur les qualifications*, see https://www.cereq.fr.

[18] For instance, interviews, after 7 years, yield a sample of around 16,000 male and female individuals in Generation 1998.

[19] The first survey, launched in 1992, has a slightly different structure.

of each employment spell, and giving the rate of employment (*i.e.*, full time, part time, etc., expressed as a percentage of full time work, between 0 and 1). In addition, we observe the wages at the moment of the interviews, after 3, 5 and 7 years. Wages are given as monthly nominal salaries, including bonuses, net of compulsory social security and medical-insurance contributions, but gross of the income tax.[20] Individuals typically leave the education system on a given month during the base year, but not all on the same month. Thus, there is some variation of the beginning month, so that young workers accumulate different amounts of potential experience during the period covered by the survey. Individuals also take a variable number of months to find a first job.[21] To sum up, we observe employment variables for a sample of individuals every month from 1998 until 2005, from 2003 to 2011 and from 2009 to 2017, but we observe wages only at some dates: at the endpoints of employment spells and at the moment of interviews.

Stacking the three *Generation* surveys of 1998, 2004 and 2010, we obtain a standard, but unbalanced panel. The panel is unbalanced for two reasons. Firstly, we do not observe the individuals of a given survey in all periods from January 1998 to December 2017. Secondly, we do not observe the wages in the middle of employment spells. For details on sample construction, see Online Appendix B15. Descriptive statistics are presented in Appendix B1.

2.3.3 Do we Observe a Devaluation of Degrees?

We defined the devaluation of a degree above as an absolute decrease in the expected real salary of workers conditional on holding the degree.[22] Relative devaluations refer to drops in the College skill premia or more generally to a decrease in the ratio of average wages conditional on two different degree categories. The main questions that we ask are simple: do we observe a devaluation of degrees over this period of 20 years, and if a devaluation did indeed occur, what are its likely causes? In particular, can it be attributed to changes in the selection of students?

For estimation, we limited ourselves to full-time wages. With the restriction

[20] This is not exactly the usual take-home pay, but note that before January 2019, the French income tax was not withheld from wages. The definition used here was therefore, for most individuals, the most easily observable and most salient expression of their income.

[21] In practice, the realization of interviews after 3, 5 or 7-years last 3 or 4 months (from October to December).

[22] Our definition of devaluation is simple and empirical. We can estimate the average real wage of a student holding a certificate of some given category after 7 years of career, in 2005, 2011 and 2017 and compare these averages at several points in time. If the average real wage decreased, we say that this particular category of degrees is devalued.

to full-time wage observations, we clearly maximize the chances of selecting individuals in relatively good health, with relatively good jobs and good pay.[23] If, given this kind of selection, we observe a devaluation, it is therefore all the more significant.

We started with a preliminary analysis of the data using standard econometric methods, including the usual panel-data *within* estimators. Our preliminary study shows a devaluation of higher-education degrees, and more specifically Master's degrees, of the order of 10%, between 1998 and 2017. Part of the devaluation is in fact due to a drop in returns to experience. We also studied the effect of the business cycle on wages, in a simple way. The Online Appendix B8 gives a presentation of these preliminary results.[24]

To summarize the preliminary analysis, we believe that the devaluation of higher-education degrees is most likely due to an excess supply of graduates. Yet, we know that there are competing explanations. The value of the degrees under scrutiny depends on (at least) two other factors: the selection of student skills and the quality of education. Both factors contribute to the graduates' human capital, and therefore to productivity and wages. It is a common contention that the quality of students enrolled in advanced programs has gone down in the recent years (this is heard in France and elsewhere). The quality of the teaching may also have decreased with time, and the two phenomena go hand in hand. At this point, with the help of standard econometric methods, it is almost impossible to decide if the selection of talents enrolled in higher education has changed in the past twenty years. To push the investigation further, we therefore propose a model of unobserved heterogeneity.

2.4 The Model

To model the beginning of careers of three cohorts of young men under unobservable heterogeneity, we assume that the distributions that we see are generated by a finite mixture of distributions, each point in the mixture being a latent, unobservable type of individual.

Let c denote the cohort of the individual with $c \in \{1998, 2004, 2010\}$. In each

[23] The advantage of doing this is to avoid possible errors in the reporting of part-time work and part-time wages, given that we do not observe the exact number of hours, and that most of full-time jobs correspond to 140 hours per month.

[24] Our preliminary analysis, in essence, is also exposed in Argan and Gary-Bobo (2021) — but the latter article is written in French.

cohort, we follow individuals across time from the moment they leave school to the moment of the survey seven years later.

We denote by t the elapsed time period (in months) since the first individuals of the first cohort left school. Note that, at $t = 1$, individuals do not have the same age, as some individuals just graduated from high school and enter the labor market while others just graduated from university. Individuals are indexed by i, with $i = 1, \ldots, N$. Let h index the highest level of education reached by the individual with $h = 1, \ldots, H$. Let $\chi_h(i)$ be the dummy that indicates whether individual i has reached education level h. Let X_{it} denote a vector of observed characteristics of i at time t. We decompose X_{it} in two subsets of variables, Z_i, the set of time-invariant variables and Ξ_{it}, the set of time-varying characteristics, so that $X_{it} = (Z_i, \Xi_{it})$.

Let W_{it} denote the observed real salary of i at time t. Let $w_{it} = \ln(W_{it})$. To obtain real wages, we deflated nominal wages, using the French consumer price index.[25] We also observe the employment rate of individual i at date t, denoted e_{it}. The latter variable takes on a finite number of values only, $e_{it} \in \{0, .3, .5, .6, .8, 1\}$; $e = 1$ represents full-time employment, and numbers between 0 and 1 measure the hours of part-time jobs as a fraction of a standard full-time job. Using the convention that $e_{it} = 0$ for all periods t such that i has not yet left the educational system, we therefore also measure effective experience, denoted x_{it}, as the cumulative hours of work, that is, for $t > 1$,

$$x_{it} = \sum_{\tau=1}^{t-1} e_{i\tau}, \tag{2.1}$$

where $x_{i1} = 0$.

We assume that individuals belong to one of a finite number of unobserved groups, called *types*. Let K be the number of latent types and let k index types. We denote $\theta_k(i)$ the dummy that indicates whether individual i is of type k.

2.4.1 Wage equation

Note that individual i's wage is not observed each month (for each t). The wage is observed at the onset and at the end of employment spells, and at the moment of the survey. Let T_i be the subset of dates t at which we observe a wage for individual i.

We can now specify the wage equation. For $t \in T_i$ and for an individual i of

[25] We used the CPI published by the National Statistical Institute, *i.e.*, INSEE. All wages ar expressed in 2013 euros.

type k, we set

$$w_{itk} = \alpha_{0k} + \sum_{c=1}^{C} \chi_c(i) \left(\delta_{0ck} + \sum_{h=1}^{H} \chi_h(i) \left(\gamma_{0chk} + \beta_{0chk} x_{it} \right) \right) + X_{it} \eta_{0k} + \varepsilon_{itk}, \quad (2.2)$$

where ε_{itk} is a normal error term with a zero mean and variance σ^2_{wk}, where $\chi_c(i)$ is a dummy indicating if the individual is in cohort c and $(\alpha_{0k}, \beta_{0chk}, \gamma_{0chk}, \delta_{0ck}, \eta_{0k})_{h=1,.,H,k=1,.,K,c=1,.,C}$ is a vector of parameters. In addition, note that x_{it} is effective experience as defined by equation 2.1. Given this, the expression for the observed wage of individual i at period t is,

$$w_{it} = \sum_{k=1}^{K} \theta_k(i) \left[\alpha_{0k} + \sum_{c=1}^{C} \chi_c(i) \left(\delta_{0ck} + \sum_{h=1}^{H} \chi_h(i) \left(\gamma_{0chk} + \beta_{0chk} x_{it} \right) \right) + X_{it} \eta_{0k} \right] + \varepsilon_{it}, \quad (2.3)$$

where $\varepsilon_{it} = \sum_k \varepsilon_{itk} \theta_k(i)$.

Note that the model is very flexible insofar as all the parameters of the wage equation are free to vary with type k; returns to education and experience vary with the cohort c and returns to experience may also depend on educational attainment h.

2.4.2 Employment equation

We model the employment level e_{it} at each date by means of an Ordered Probit model. Recall that e_{it} takes on discrete values between 0 and 1 that measure individual i's rate of employment in period t. Let G be the number of levels of employment, denoted $\mathbf{e}_g$, with $g = 1, .., G$ and $1 \geq \mathbf{e}_{g+1} > \mathbf{e}_g \geq 0$. We define,

$$P_k(e_{it}|X_i, x_{it}, h_i) = \Pr(e_{it} = \mathbf{e}_g \,|\, X_{it}, x_{it}, h_i, k) = \Pr\left[\mathbf{c}_{gk} \leq \rho_{itk} + \zeta_{itk} \leq \mathbf{c}_{g+1,k} \right], \quad (2.4)$$

where

$$\rho_{itk} = \sum_{c=1}^{C} \chi_c(i) \left(\delta_{1ck} + \beta_{1ck} x_{it} + \sum_{h=1}^{H} \gamma_{1chk} \chi_h(i) \right) + X_{it} \eta_{1k}, \quad (2.5)$$

where the $\mathbf{c}_{gk}$ are the thresholds (*i.e.*, cuts) of the Ordered Probit, and $\mathbf{c}_{0k} = -\infty$. The ζ_{itk} are independent random variables with a standard normal distribution and $(\beta_{1ck}, \gamma_{1chk}, \delta_{1ck}, \eta_{1k})_{h=1,.,H,k=1,.,K,c=1,2,3}$ is a vector of parameters to estimate. Remark that all parameters are free to vary with k.

We denote E_i the subset of dates t at which e_{it} is observed. This model is estimated mainly thanks to observations e_{it} at the beginning and the end of each

employment spell of individual i. In addition, there are some observations in the middle of a spell. Typically, this happens when, at the end of the survey period, an individual is currently employed, and these observations correspond to truncated spells. Typically, at a date t corresponding to the beginning of an employment spell, the employment rate e_{it} jumps to 1, or a positive value smaller than 1 in the case of a part-time job. At a date t corresponding to the last period of a full-employment spell, we observe $e_{i,t+1} = 0$ if i becomes unemployed, or $0 < e_{i,t+1} \leq 1$ if i changes for a part-time job.

This model is clearly a kind of reduced form, but it is rich and flexible enough to capture the possibility that probabilities of finding a job at any t depend on accumulated experience, degrees, the cohort, and the type.

2.4.3 Education equation

Finally, we model the level of education with the help of a multinomial logit model. This approach provides a simple way of modelling individual investment in education. We denote Λ the probability of choosing education h, that is,

$$\Lambda_k(h|Z_i) = \Pr(h_i = h \,|\, Z_i, k) = \Pr\left[\mathbf{u}_{ihk} = \max_{j\in\{1,\ldots,H\}} \left(\mathbf{u}_{ijk}\right)\right], \tag{2.6}$$

where the utility u_{ihk} of an individual i of type k choosing education level h is defined as $\mathbf{u}_{ihk} = \mathbf{v}_{ihk} + \xi_{ihk}$, with

$$\mathbf{v}_{ihk} = \alpha_{2hk} + \sum_{c=1}^{C} \delta_{2chk}\chi_c(i) + Z_i\eta_{2hk}, \tag{2.7}$$

where ξ_{ihk} is a random variable that follows a Gumbel distribution (*i.e.*, Type 1 extreme-value distribution) and where we want to estimate the following vector of parameters: $(\alpha_{2hk}, \delta_{2chk}, \eta_{2hk})$ where $h = 1,\ldots,H$, $k = 1,\ldots,K$, and $c = 1,\ldots,C$.

Clearly, this model is again a reduced form. The description of education choices is static. In addition, the model has a "triangular" structure because degrees explain experience and degrees and experience explain wages. In other words, wages do not appear in the choice equations. But as discussed in the literature, the ex ante wage expectations of individuals should in principle appear in the choice equations, instead of the ex post, effectively observed wages of each individual.[26]

[26] On this theme, see Beffy et al. (2012) and Arcidiacono et al. (2020). On *ex ante* returns to schooling, on the separation of what a student can forecast at the time of educational decisions, based

This would require a model of wage expectations depending on the latent types — a possible extension of our approach. Since education will depend on the latent groups, we can say that types capture differences in expectations in a rudimentary way. The multinomial choice equation may be viewed as an auxiliary part of the model, yet, it permits us to estimate choice probabilities that depend on the latent types.

2.4.4 Identification

We estimate the model by maximization of the log likelihood. We typically use the sequential EM algorithm to obtain preliminary estimates, and then use a standard ML algorithm. The model's likelihood is derived in Appendix B2.

The maximum likelihood method provides us with estimated values and standard deviations for all parameters, $(\alpha, \beta, \gamma, \delta, \eta, \sigma, \mathbf{c})$ and the *prior* probabilities of types p_k. We present here the results obtained when we fix $K = 3$. We discuss the choice of the number of types, using information and entropy criteria, in Online Appendix B14. An important output of the estimation algorithm is the *posterior* probability that individual i is of type k, that is,

$$p_{ik} = \Pr(k|X_i, y_i). \tag{2.8}$$

The probability p_{ik}, can be expressed with the help of Bayes' rule and the likelihood, as indicated in Appendix B2.

Identification and Nonparametric Identification. Our main identifying assumption is that wage observations (and employment rates) are independent conditional on accumulated experience, observable characteristics (degrees) and latent types. Parametric identification of the wage equation is obtained under standard conditions (see McLachlan and Peel (2000)). The ordered probit and the multinomial logit would be parametrically identified in the case of a single type. In addition, a static discrete choice model, if estimated separately, does not permit the identification of latent choices. We will come back to this point below.

We first discuss the identification of a wage equation with a latent structure. The discussion on the possibility of *nonparametric identification* can be based on

on private information, from the risk in future wages, *i.e.*, the separation of risk from heterogeneity in the observed distribution of wages, there is an important literature; see Cunha and Heckman (2007), Carneiro et al. (2003), Cunha et al. (2005).

the results of Allman et al. (2009). In a nutshell, our wage equation alone would allow us to identify a latent type structure and its parameters nonparametrically, up to a relabeling of types, *i.e.*, we would obtain, for given K, the probabilities of types p_k and the conditional c.d.fs $G_t(w|k)$ of wages w at time t. So, in principle, we could get rid of the normality assumption and still estimate the wage model with a set of latent types and their associated probabilities. More precisely, to achieve full nonparametric identification, according to the theorems of Allman *et al.* (2009), we need three groups of variables that are independent conditional on the latent types, plus a condition that the conditional distributions $G(.|k)$, $k = 1,\ldots,K$, are linearly independent.[27] The latter condition is reasonable if types are really different. So, the main problem is to find three conditionally independent random measures of types: we now show that the three measures are at hand.

We can apply the general theorems if we also condition with respect to observable characteristics. The employment rate profile of any individual, and therefore this individual's profile of accumulated experience, can be described by a finite number of states or cells, since employment rates e_{it} are discretized. Other observed characteristics such as the educational achievement h and the family-background variables are typically dummy variables (if a control is continuous, it can be discretized). It follows that we can bin the entire population in a finite number of cells. Given our assumption on wages (and the wage equation above), *in each of these cells*, and *conditional on the latent type*, wage observations made at different dates t are independent. In our panel, at least three different values of w are available for each i. Now, let K be the number of types. For each k, we identify in each cell X a probability $p(k|X)$ and an array of distributions $G_t(w|k,X)$. Given that we know the distribution of observable variables $\phi(X)$, we easily derive $p_k = \sum_X \phi(X)p(k|X)$, etc. It follows that a latent type structure can be nonparametrically identified from the distribution of wages.

A more difficult problem is to nonparametrically identify a finite latent structure for the joint distribution of wages, employment rates and educational choices. The theorems of Allman *et al.* (2009) cannot be applied because education determines employment and wages, and because employment (in fact, experience) determines wages: the three variables cannot be assumed independent conditional on the latent types.

[27] See particularly Theorem 8 in Allman *et al.* (2009). On this topic, further results are proved and estimation methods are provided in Bonhomme et al. (2016).

The literature on the identification of dynamic discrete choice models[28] provides us with some tools that can be applied to the study of our model. Our Ordered Probit model, used to predict the employment rate at each t, which is a specific discrete choice model, is nonparametrically identified using the results of Kasahara and Shimotsu (2009). In the latter paper, the key features permitting nonparametric identification of a finite mixture are: *(i)* the observation of individual choices during a sufficiently large number of periods (*i.e.*, the length of the panel), *(ii)*, the number of different values that time-varying control variables can take; and *(iii)*, the fact that latent types react differently to changes in the control variables. Our panel is sufficiently long; the accumulated experience varies with time in many possible ways; it is reasonable to assume that each type reacts differently to changes in effective experience: nonparametric identification is at hand.

Finally, the education choice model is static and it follows that a finite mixture of multinomial choice models cannot be identified in isolation. Yet, if we fix the number of types and know their probabilities, we can obtain the choice model for each type simply by means of a weighted likelihood-maximization algorithm, as in the M-step of an EM algorithm, in spite of the fact that the model is static. The finite mixture of multinomial choice models can therefore be identified jointly with the wage equation, since the latter provides the type probabilities that are needed to estimate its parameters. In other words, the wage model provides an auxiliary equation for the finite mixture of Multinomial Logits. To conlude this discussion, it is possible to obtain a nonparametric identification result for the complete model, but it is a nontrivial problem to prove such a result rigorously, and this problem is beyond the ambitions of the present article.

2.5 Policy-relevant parameters; ATEs and ATTs

We will use the model, the estimated values of parameters and the posterior probabilities of types for each individual, denoted p_{ik}, to compute policy-relevant parameters. In particular, we can study ATTs, ATEs and the effect of unobserved selection on various outcomes. It is also possible to study the heterogeneity of treatment effects and we can compute ATEs and ATTs conditional on type k.

[28] See also Magnac and Thesmar (2002), Hall and Zhou (2003).

2.5.1 Policy-relevant parameters; ATEs and ATTs: Method

Let $y_t(z)$ denote the potential value of any outcome, at time t, for individuals with observable characteristics z.[29] We first define an *average treatment effect conditional on type k* and education level h at time t, denoted $ATE(h,k,t)$. Let $h=0$ denote the level of individuals without any degree (high-school dropouts): these individuals are our reference point. This conditional treatment effect is defined as follows,

$$ATE(h,k,t) = \mathbb{E}[y_t(h)|k] - \mathbb{E}[y_t(0)|k]. \tag{2.9}$$

The (unconditional) average treatment effect at time t, for individuals with level h is then defined as follows,

$$ATE(h,t) = \sum_k p_k ATE(h,k,t), \tag{2.10}$$

where the p_k are the prior probabilities of types defined above.

For any vector of observable characteristics z, let $\chi_z(i)=1$ if and only if $z_i = z$ and $\chi_z(i)=0$ otherwise. We use the observations y_{it} of the outcome for individuals i. To estimate $\mathbb{E}[y_t(h)|k]$ we use the statistic,

$$\hat{\mathbb{E}}[y_t(h)|k] = \frac{\sum_i y_{it}\hat{p}_{ik}\chi_h(i)}{\sum_i \hat{p}_{ik}\chi_h(i)} = \frac{\sum\limits_{\{i|h_i=h\}} y_{it}\hat{p}_{ik}}{\sum\limits_{\{i|h_i=h\}} \hat{p}_{ik}}, \tag{2.11}$$

where $\hat{p}_{ik}$ is the estimated posterior probability that i belongs to group k, computed by Bayes's law as indicated above by expression (2.8). Basically (2.11) is an estimation of $\mathbb{E}(y|h,k)$, using the sample. In a similar fashion, we define,

$$\widehat{ATE}(h,t) = \sum_k \hat{p}_k \widehat{ATE}(h,k,t), \tag{2.12}$$

where $\hat{p}_k$ is the estimated prior probability, and where,

$$\widehat{ATE}(h,k,t) = \hat{\mathbb{E}}[y_t(h)|k] - \hat{\mathbb{E}}[y_t(0)|k]. \tag{2.13}$$

[29] For instance, the potential average wage of an individual with degree h and t months of potential experience, *i.e.*, $w_t(h)$, is an outcome of interest, as well as the average wage of an individual with characteristic z in cohort c, that we can also denote $w_c(z)$ (with a slight abuse of notation). In a similar fashion, we define the employment rate $e_t(z)$; the accumulated level of effective experience $x_t(z)$, etc.

Now, to compute the ATT (average effect of treatment on the treated), we have,

$$\mathbb{E}[y_t(h)|h',k] = \mathbb{E}[y_t(h)|k] \quad \text{for all } (h',h). \tag{2.14}$$

This is equivalent to the familiar *conditional independence assumption* of the treatment-effects literature, except that conditioning is with respect to the unobservable type k, and education h is the treatment here. In other words, the expected counterfactual (or potential) outcome of a type k with degree h', if instead of h' they had chosen a degree h, is just the mean outcome of individuals with degree h, knowing type k. Under this assumption, we have

$$ATT(h,k,t) = ATE(h,k,t), \tag{2.15}$$

and it is easy to show that,

$$ATT(h,t) = \sum_k p(k|h) ATE(h,k,t), \quad \text{where} \quad p(k|h) = \frac{p(h,k)}{p(h)}. \tag{2.16}$$

Now, obviously, to estimate ATT, we use $\widehat{ATE}(h,k,t)$ and the estimated conditional probability $\hat{p}(k|h)$ which is itself the ratio of[30]

$$\hat{p}(h,k) = \frac{1}{N}\sum_i \hat{p}_{ik}\chi_h(i) \tag{2.17}$$

and

$$\hat{p}(h) = \sum_k \hat{p}(h,k) = \frac{1}{N}\sum_i \chi_h(i). \tag{2.18}$$

Finally, $\widehat{ATT}$ is just obtained by putting hats on p and ATE in equation 2.16.

With the help of posterior probabilities, we can estimate the probability of choosing h, knowing unobservable type k and observable characteristic z as follows,

$$\hat{p}(h|k,z) = \frac{\hat{p}(h,k,z)}{\hat{p}(k,z)} = \frac{\sum_i \hat{p}_{ik}\chi_{hz}(i)}{\sum_i \hat{p}_{ik}\chi_z(i)}, \tag{2.19}$$

where $\chi_{hz}(i) = 1$ iff $z_i = z$ and $h_i = h$ and $\chi_{hz}(i) = 0$ otherwise.

In the case of wages, we do not observe w_{it} at every t and for every i but we can take averages over several periods, if needed, and the definitions would be changed accordingly, in an obvious manner. For instance, we will consider averages over

[30] It is easy to check that $\sum_h \sum_k \hat{p}(h,k) = 1$.

all the periods t of a given cohort c.

2.5.2 Estimation of the discounted sum of earnings. Simulations

We can estimate the "human capital", *i.e.*, the discounted sum of earnings of an individual i with observable characteristics Z_i and type k, using the estimated model. This outcome is interesting to compare types, because it summarizes all the differences in wages (returns to education and experience), employment rates and educational achievement. To this end, we simulated a fictitious sequence of employment, experience and wages for each individual i in the sample. Then, using weights p_{ik}, we averaged the expected-discounted fictitious sequence of earnings of each i. We provide details on the method of simulation in Appendix B3.

2.6 Results

We can now present our estimation results. We start with the distribution of types. Next we compute the ATTs and ATEs of education and the simulated discounted earnings, type by type, and overall. Then, we present the ML estimates of the model's other parameters and we discuss returns to education, returns to experience and educational choices, conditional on unobserved types. This leads us to propose an interpretation of the three types that we find: the types are clearly different, with a clear hierarchy.

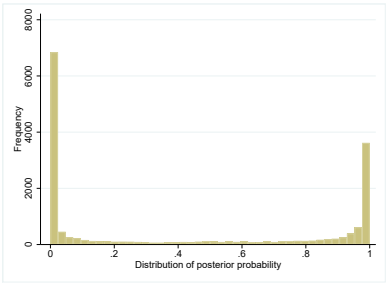
To estimate the model, in addition to the variables discussed in Section 3, we use the following list of variables X_t: the student's location in geographical space, indicated by dummies (Urban, Peri-Urban and Rural), the father's occupation (the father-is-a-professional dummy); and the macroeconomic unemployment rate. Further explanations about controls are given below.

2.6.1 Probability of types *k=1,2,3*

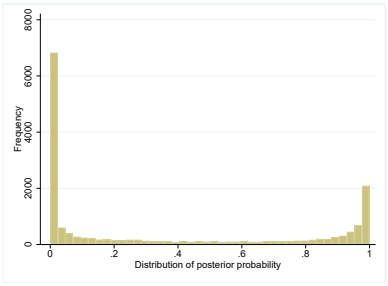
Table 2.1 presents the estimated probabilities of types when $K = 3$: in our sample 42% of young men are of type 1, 36% of type 2 and 22% of type 3. The type frequencies are very precisely estimated. Before we provide an interpretation of these types —in other words, who do these types represent? — it is important to check if these types generate a good classification (*i.e.*, a near partition) of the population. The quality of classification is said to be good if each individual i belongs to a given group k with a sufficiently high probability, say, ideally, with

Figure 2.2: Empirical distribution of posterior type probabilities

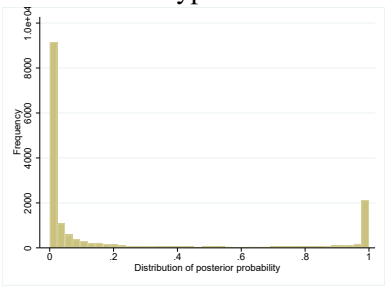
Type 1



Type 2



Type 3



$p_{ik} \simeq 1$ for some k. It may happen that a minority of individuals remains hard to categorize, and for these, we would find $p_{ik} \simeq 1/K$, or alternatively, they sometimes belong to a subset of $K' < K$ types with a high probability. Visual inspection of the histograms of the estimated values p_{ik} for each k will immediately show if the classification is fuzzy. Figure 2.2 shows that the classification is in fact very good. Most individuals have values of p_{ik} close to 0 or 1.

Table 2.1: Estimated probability of types

Type	1	2	3
Probability	0.42	0.36	0.22
Standard error	(.006)	(.006)	-

Table 2.2: Distribution of types by cohort

Type	1	2	3
1998	0.40	0.36	0.23
2004	0.42	0.37	0.21
2010	0.45	0.33	0.22

Given our results, it seems that the types are not simply fictitious disembodied categories used to fit the distribution of employment and wages: they are likely to correspond to real people. It remains to understand which observable characteris-

tics help recognizing a given type.

As explained above, our ML estimation results permit us to examine the distribution of types conditional on any characteristic or set of characteristics. It is sufficient to compute the arithmetic average of posterior probabilities p_{ik} in the subset of individuals i sharing the given characteristic(s). The distribution of types by cohort is presented in Table 2.2. It appears that this distribution is very stable across cohorts. None of the types is negligible — all the prior probabilities (or frequencies) are above 20%. In the coming subsections, we will see how individual wages and employment rates depend on types. The multinomial, discrete-choice part of our model gives a more detailed account of individual education decisions.

2.6.2 ATEs, ATTs and Discounted Earnings: Results

We now present the values of ATEs and ATTs, as well as the simulated values of discounted earnings by type.

ATEs and ATTs. Changes in the Selection of Students

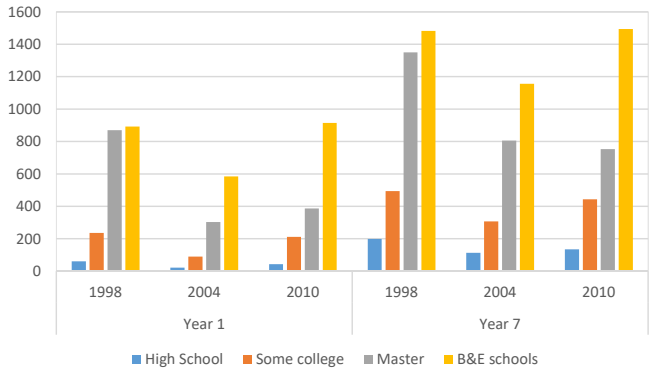
The ATE and ATT parameters are defined above, in subsection 5.1., by Equations 2.12, 2.13 and 2.16, 2.17, 2.18. Figure 2.3 gives: a) the ATE; b) the ATT; and c) the percentage variation (ATT − ATE)/ATE for full-time wages. The left-hand part of each panel of Fig. 2.3 displays the results for wages observed during the first year while the right-hand part of each panel displays the results obtained with wages observed during the seventh year of career. The reference education $h = 0$ is the category of 'less than high-school degree' (including the dropouts).

The first striking result is that ATE is consistently larger than ATT for the high-school graduates, *i.e.*, the students for which the high-school degree is the highest degree. This confirms our intuition that this education level does not select the most productive students.

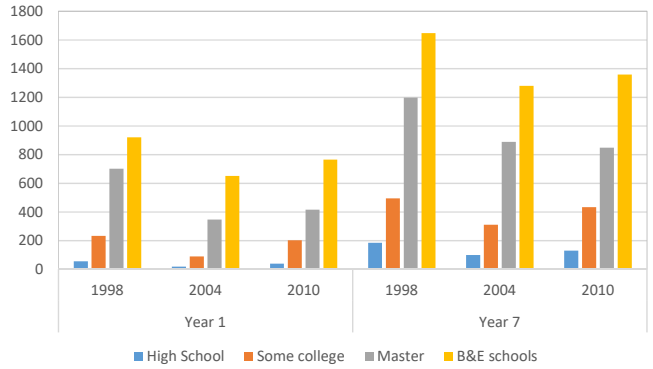
The second striking result that is visible is the drop in the ATE of Master degrees: between 2017 and 1998, the absolute variation ΔATE after 7 years of career is around -600 euros per month. The absolute variation ΔATT, during the same period and for the same degrees is around -400 euros (per month), a smaller drop. The bottom panel of Fig. 2.3 shows that in 1998, we had ATE¿ATT for the Master program graduates. These graduates were therefore less well selected than the average population, but this difference reversed in the later years. Indeed, in the 2004 and 2010 cohorts, panel *c* of Fig. 2.3 clearly shows that the selection of Mas-

Figure 2.3: ATE and ATT when education is the treatment and full-time wages are the outcome

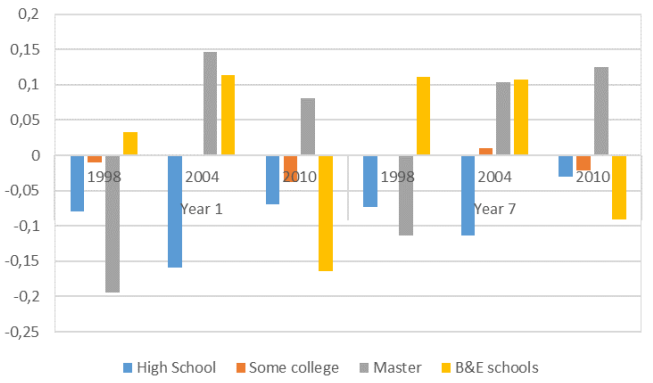
a. ATE



b. ATT



c. (ATT − ATE)/ATE



ter graduates has improved. The reversal is particularly visible after 7 years. This result is surprising, because we expected a selection of lesser quality students in these programs, due to the sharp increase in enrollment. A consequence of this observation is that an 'excess supply' of graduates could be the main explanation for degree devaluation, because it is not the result of adverse selection.

Next, the situation is more complicated than it may seem at first glance, because the evolution of selection is exactly the opposite for business-school and engineering-school graduates. The difference ATT-ATE clearly decreased between the 1998 and 2010 cohorts. There is a lot of evidence about the constant growth of business schools in France. These schools have been growing since the 1970s and new schools opened. The growth of aggregate enrollment accelerated in the recent years, in spite of increasing tuition fees. The growth was possible only at the cost of less selectivity.[31] The interpretation of our results is therefore straightforward. In sharp contrast, University degrees, in spite of the growth of enrollment, have markedly improved selection at the master's level. In fact, if we put business and engineering schools and the doctorates aside, the master's degree has become the only really selective instrument of French universities.

Simulations of the Model. Discounted Earnings Conditional on Type k

We simulate sequences of employment rates and wage rates $(\tilde{e}_{itk}, \tilde{w}_{itk})$. This allows us to compute the discounted expected earnings during the periods $t \in \{1, \ldots, T\}$. We choose a discount factor $\delta = .99$ (per month) and for every (i,k), we compute,

$$\tilde{W}_{ik} = \frac{(1-\delta)}{(1-\delta^T)} \sum_{t=1}^{T} \delta^{t-1} \tilde{e}_{itk} \exp(\tilde{w}_{itk}).$$

$\tilde{W}_{ik}$ is a weighted average and has the dimension of monthly earnings. Then, we compute the weighted arithmetic mean, using the estimated probabilities p_{ik}. For each type k, we compute,

$$H_k = \frac{\sum_{i=1}^{N} \tilde{W}_{ik} \hat{p}_{ik}}{\sum_{i=1}^{N} \hat{p}_{ik}}.$$

See Appendix B3 for a detailed description of simulations. The simulations are based on the full estimated model. The value of H_k can be computed in subsamples,

[31] Indeed, it is well-known that the French business schools developed teaching programs, like the "bachelors" that cannot be compared with the traditional programs of the French "grande école". Note that the individuals in the business-school category here, as already mentioned, have completed 5 years of study after high-school graduation, and can be compared with Master's graduates.

conditional on c or h or both. The results are given by Table 2.3. The figures are rather low for Type 1. This is due, not only to smaller monthly wages, but also to a lot of unemployment. In addition, we compute these values conditional on the cohort c, denoted $H_k(c)$. We see a clear hierarchy of types. Type 1 did

Table 2.3: Discounted earnings by type and cohort, *i.e.*, $H_k(c)$

Type	1	2	3
All cohorts	746	1215	1329
1998	740	1173	1091
2004	745	1252	1393
2010	758	1246	1334

not experience a devaluation (a drop of $H_k(c)$ with c) but this devaluation hit the other two types, between the 2004 and 2010 cohorts. The discounted values H_k are a synthetic indicator summarizing the effects of returns to degrees, returns to experience and unemployment. We see that Type 3 has lost 4.2% between 2004 and 2010.

Table 2.4: Average discounted earnings by type, degree and cohort, *i.e.*, $H_k(c,h)$

Cohort	1998			2004			2010		
Type	1	2	3	1	2	3	1	2	3
Less than High School	652	1017	819	635	994	934	510	865	930
High-School Degree	711	1050	1094	723	1113	1145	664	1072	952
Some College and Bachelors	802	1239	1382	801	1273	1507	772	1288	1496
Masters (M2)	1412	2362	1121	1017	1628	1879	837	1357	1974
Bus. & Eng. Schools	1327	1924	1861	1151	1786	2215	1874	2411	1443

Table 2.4 shows the discounted earnings by type, degree and cohort, *i.e.*, $H_k(c,h)$. It is striking to see that, in terms of discounted earnings, devaluation took place for the less-than-high-school and the high-school degrees of Type 1 and 2. This is due to worse employment conditions because the wage rates increased, mainly as a consequence of an increased minimum wage. In the 'Some College and Bachelors' category, the devaluation hits Type 1 only. The devaluation of Masters is confirmed for Types 1 and 2, but not for Type 3: we give more details on this result below, in subsection 6.3. The interpretation of results obtained for business and engineering schools is more delicate: macroeconomic conditions probably play a role in explaining the unstable performances of Type 3 (but Type 3 is characterized by relatively less stable jobs, as compared to Type 2, as we will see in subsection 6.3 below).

Do we see selectivity changes in the sub-populations of graduates? Table 2.5,

Table 2.5: Discounted earnings by degree and cohort, *i.e.*, $H(h,c)$

Cohort	Degree	Actual	Counterfactual	Difference	Percent Variation
1998	Less than High School	813	824	−11	−0.1%
	High-School Degree	916	924	−8	−0.8%
	Some College and Bachelors	1105	1097	+9	+0.8%
	M2	1589	1689	−100	−6.2%
	Bus. & Eng. Schools	1782	1669	+113	+6.3%
2004	Less than High School	806	831	−25	−3.1%
	High-School Degree	932	956	−24	−2.6%
	Some College and Bachelors	1146	1125	+21	+1.8%
	M2	1530	1426	+105	+6.8%
	Bus. & Eng. Schools	1769	1611	+158	+8.9%
2010	Less than High School	691	720	−30	−4.3%
	High-School Degree	828	862	−34	−4.1%
	Some College and Bachelors	1121	1103	+17	+1.5%
	M2	1414	1261	+153	+10.8%
	Bus. & Eng. Schools	1862	1954	−92	−4.9%

in column 3 (*i.e.*, 'Actual') gives the average value of $H_k(h,c)$, weighted by the conditional probabilities $p(k|h,c)$, while column 4 (*i.e.*, 'Counterfactual') gives the average of $H_k(h,c)$ weighted by probabilities $p(k|c)$. The fifth column of Table 2.5 gives the difference *Actual* − *Counterfactual*. This difference measures the extent to which individuals are positively or negatively selected at various educational levels. The less-than-high-school-degree and high-school-degree holders earn on average less than if this population had the distribution of types of the whole population. The figures in the Selection column are negative in the three cohorts, but the difference between Actual and Counterfactual is small. In contrast, the situation of M2 degree holders has changed with time. Selection was clearly negative in 1998 (theses graduates seem less able than the general population), but the selection becomes positive in the 2004 and 2010 cohorts. The number of students enrolled in master programs has increased, but in fact, these university degrees have selected students that seem better than the average in a certain sense: they tend to have a higher type.

Next, the discounted earnings of engineering and business-school graduates has followed a completely different path: it seems that the quality of the selection of schools has deteriorated with time. These results confirm the findings obtained above with ATE and ATT when observed wages are the outcome and education is the treatment.

2.6.3 Parameters estimates

We now consider in turn the ML estimation results of the three building blocks of our model: the wage equation, the employment equation; the education choice equation. The results again show a clear hierarchy of types.

Wage equation

Complete ML estimates of the wage equation are presented in Appendix B4, in Tables B.2, B.3 and B.4. Table B.2 presents the wage returns to experience by cohort, type and education level (β_{0chk}). Table B.3 presents the returns to education by cohort and type (γ_{0chk}). Table B.4 presents the other parameters of the wage equation ($\alpha_{0k}, \delta_{0ck}, \eta_{0k}$). A glance at Figures 2.4 and 2.6 will show the main insights that can be drawn from the wage equation.

Returns to experience. Figure 2.4 shows the returns to experience in the three cohorts. Remember that these returns are average percentages of wage growth by month. The colors (and intervals) are distinguishing the 3 types, while each group of three intervals corresponds to a level of educational achievement. The most striking phenomena are, firstly, that type 3 (yellow intervals) has markedly higher returns to experience than the other types; secondly, that returns to experience typically increase with the level of educational achievement[32]; thirdly, returns to experience have tended to decrease with time, between 1998 and 2017, in particular for type 3. This is shown on Fig. 2.5 for the Master's degree holders. We see an "erosion" of the returns to experience.[33] The fall in returns to experience has been particularly important for the type 3 individuals who graduated from business and engineering schools. An exception is the return to type-2, business-school graduates, that has strongly increased in the most recent cohort.

Returns to education. We now turn to estimated returns to education or returns to degrees. These returns can also be viewed as returns at zero experience determining average starting salaries. The three panels of Figure 2.6 give the returns to education for the three cohorts, the three types (represented by three intervals with different colors) and the 5 education levels.

[32] Business schools are an exception to these rules.

[33] In particular, the drop appears in the 2004 cohort; it then stabilized in the 2010 cohort for types 2 and 3, or it slightly increased again, without catching up the 1998 level for type 1.

Figure 2.4: Monthly Returns to Experience by Type, Educational Attainment and Cohort

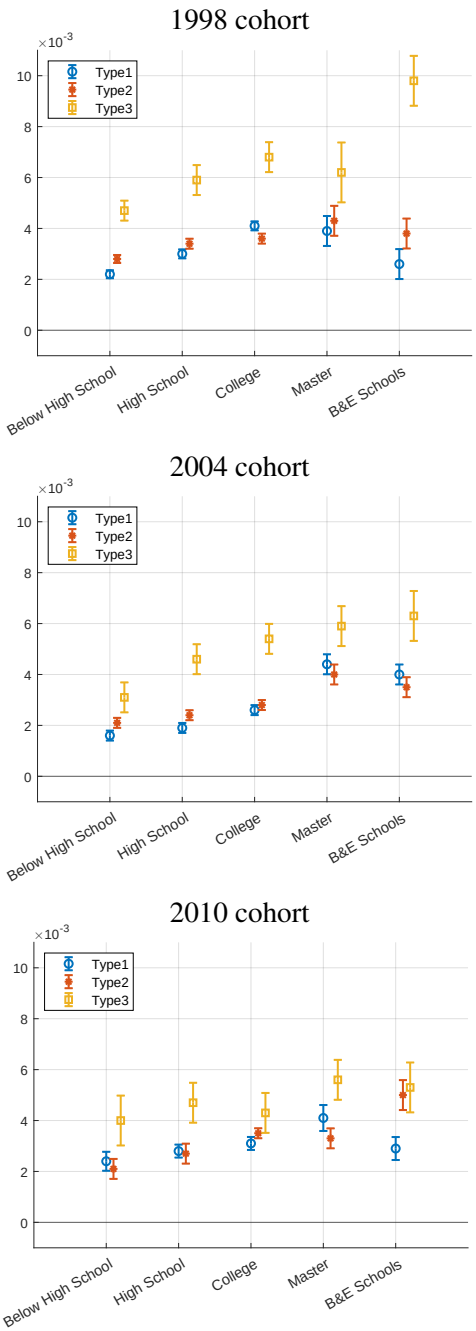
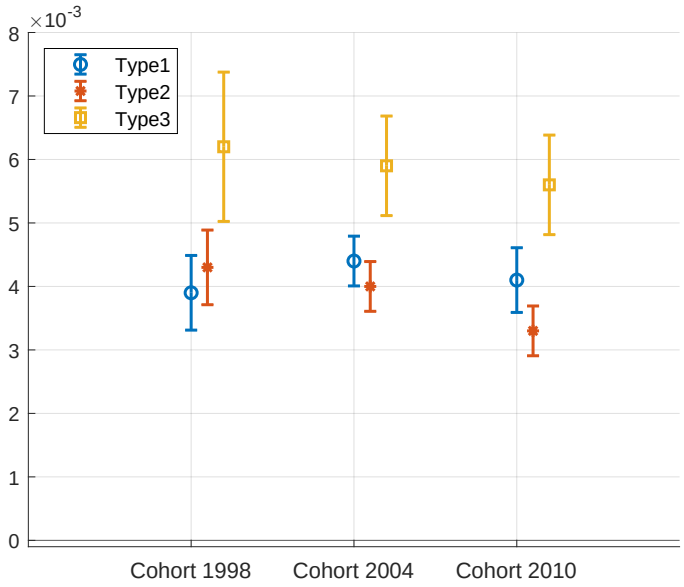


Figure 2.5: Evolution of Returns to Experience by Type: Masters



The education levels (5 groups of three bars) are clearly ranked (following the common, and expected hierarchy). In the beginning, in the 1998 cohort, type 2 gives the impression of dominating the highest educational levels, but in the 2004 cohort, we see a clear and consistent hierarchy of types: type 3 is simply the best everywhere; type 1 has the smallest returns and type 2 has median returns everywhere. The 2010 cohort confirms the hierarchy of types (with the exception of business schools).

Do we observe a devaluation of some degrees? Figure 2.7 now groups types by cohort on the same picture, to appreciate the possible devaluation of returns to degrees, in the case of Masters. Note that, on this picture, different colors now represent different cohorts. It is very striking that the average wages, conditional on type and a Master's degree have been devalued for types 1 and 2, but not for type 3. Devaluation of Masters' degrees is confirmed, but it is heterogeneous. If we compute the weighted variation of log-wages from the 1998 cohort to the 2010 cohort, using the type frequencies of Table 2.6 as weights, we find a drop of $.0747 = \Delta w \simeq 7.676 - 7.602$, and $e^{-.0747} - 1 \simeq -.072$, that is, a 7.2% devaluation of Masters' degrees. This corresponds to the result that can be obtained with a simple regression of log-wages on cohort and degree dummies (see Table B.8 in Online Appendix B8). A similar computation would show that there is no

Figure 2.6: Returns to Education by Type, Educational Attainment and Cohort

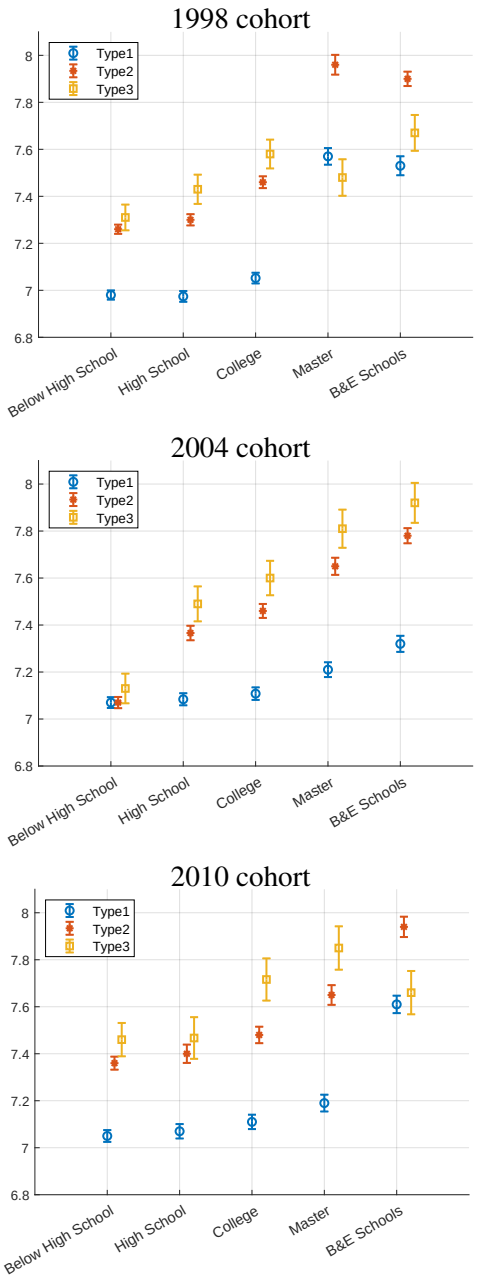
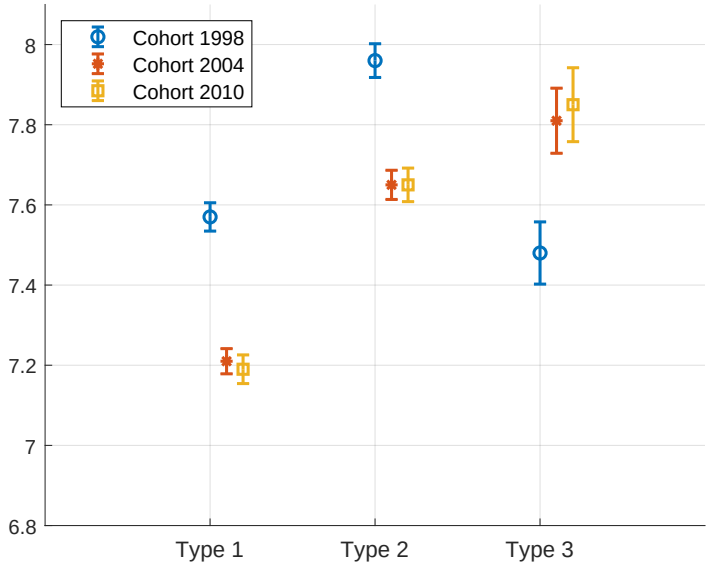


Figure 2.7: Devaluation of Master's Degrees 1998-2017

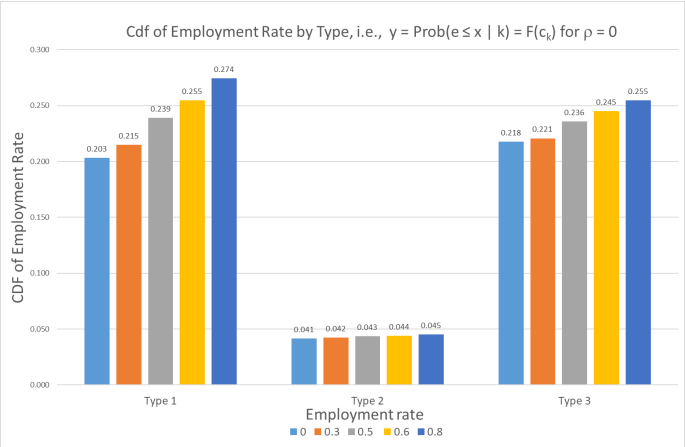


devaluation for the *Some College and Bachelors'* level. Yet, we observe a *relative* devaluation of this level, which is directly visible from the estimated coefficients displayed by Table B.3. There is a relative devaluation of Bachelors with respect to the Less-than-High-School-Degree level. As already noted, this relative devaluation is mainly due to the fact that minimum-wage regulations protect the real value of wages from depreciation at the lowest educational level. In contrast, the devaluation of Master graduates' average real wages is *absolute*.

Impact of some control variables. Table B.4, in Appendix B4, provides the estimated coefficients of some important control variables. We control the wage equation for the macroeconomic unemployment rate. This is a way of controlling for the impact of the business cycle on wages. The impact on type 1 is not very significant. This is probably due to the fact that type 1 tends to reach education levels at which wages are protected by the minimum wage legislation. But the impact of overall unemployment is clearly negative for types 2 and 3, as expected: we find mildly procyclical real wages. Secondly, we find a significant and positive effect of the *father is a professional* dummy. This latter effect is much stronger for type 3 (five times more than the effect on type 1). This dummy indicates individuals whose father's occupation requires higher-education degrees: executives, doctors, lawyers, engineers, teachers, etc. The reference individual belongs to the 1998 cohort, lives in urban areas, has a father which is not a "professional" in the above

sense. Indications of geographic origin are significant too: the rural and peri-urban individuals earn (slightly) smaller wages.[34]

Figure 2.8: Employment Rates by Type: Analysis of Ordered Probit Cuts



Employment equation

Estimates of the Ordered-Probit parameters are presented in Table B.6, Online Appendix B6. The Ordered Probit shows a striking feature of type 2 individuals. This is visible if we look at the Ordered-Probit cuts. Figure 2.8 gives a representation of these cuts. To be more precise, the table gives $\Pr(e \leq x|k) = F(\mathbf{c}_k)$, where $x \in \{0, .3, .5, .6, .8\}$, F is the standard normal c.d.f and $\mathbf{c}_k$ is the corresponding Ordered-Probit cut. In other words, we consider an individual with all controls set equal to 0 — hence we have $\rho = 0$ —, and conditional on type k, we compute the cumulative probabilities that this individual has an employment rate e smaller than x. Figure 2.8 very clearly shows that type 2 has a very small probability of unemployment (around 4%) and a high probability of full employment of 95.5%. In contrast type 1 and type 3 have, respectively, a 72.6% and a 74.5% probability of being fully employed when $X = 0$. These results give the impression that type 2 finds a job quickly and stays in this job: the matching of type 2s with employers seems very stable as compared to that of the other types. As a counterpart, these individuals obtain smaller wages at the start and, as time passes, obtain smaller

[34]The Peri-urban is a heterogeneous category including neither purely urban nor purely rural individuals: it typically includes suburban and smalltown France. Note that, unlike in America, the French suburban individual generally does not have a well-to-do background. The urban individual is more likely to come from a privileged background.

pay raises than type-3 individuals. Online Appendix B12 gives further details on the Ordered Probit and the reason for the observed differences between Type 2 and the other types in terms of employment. In particular, we study the possibility that Type 2 has a preference for the public sector.

Education choices

Table 2.6 gives the empirical values of conditional probabilities $\hat{p}(k|h,c)$, using the estimated posterior probabilities; this table shows that type 1 is more prevalent among individuals with the shortest education, in particular in the most recent cohorts. In the years 2010-17 (*i.e.*, the third cohort), more than 50% of those who did not go to college are type-1 individuals. In contrast, in the same category, we find around 30% of type-2 individuals and less than 20% belong to type 3. The distribution was closer to uniform in 1998. In 2010, there are less type-3 students in the *some college* category and a greater proportion of them in the business and engineering schools, as compared to 1998. We also see that 71% of the individuals with a Master degree are members of type 2 or type 3 in 2010, in contrast with 62% of the same categories in 1998. Also in contrast, the distribution of types among students who graduated from schools (*i.e.*, business and engineering schools) is closer to uniform in 2010 as compared to 1998 and 2004, where a large majority were members of type 2. To sum up, we see that the mix of types has changed, conditional on degrees.

Table 2.6 shows that the sorting of students has increased with time in universities and for those with an attainment below or equal to high-school graduation. Business and Engineering schools are an exception since sorting has decreased, schools admitting more members of types 1 and 3.

Table 2.6: Mix of types by education level and cohort

	$p(k\|h,c)$								
	1998 cohort			2004 cohort			2010 cohort		
Probability of type ...	1	2	3	1	2	3	1	2	3
Less than High-school Degree	0.43	0.33	0.25	0.50	0.33	0.18	0.52	0.32	0.16
High-school Degree	0.42	0.37	0.21	0.48	0.34	0.18	0.54	0.25	0.21
Some College and Bachelors	0.38	0.40	0.22	0.38	0.38	0.23	0.40	0.41	0.19
Masters	0.38	0.29	0.33	0.29	0.39	0.32	0.29	0.38	0.33
Bus. Engin. School Degrees	0.21	0.51	0.28	0.19	0.57	0.24	0.37	0.27	0.36

Table 2.7 provides a different point of view on the same reality and displays conditional probabilities of choosing level h, given the type k and cohort c, that

Table 2.7: Probability of reaching an education level given the type and cohort

	$p(h\|k,c)$								
Conditional on cohort ...	1998			2004			2010		
and conditional on type ...	1	2	3	1	2	3	1	2	3
Less than High-school Degree	0.43	0.37	0.43	0.34	0.26	0.24	0.29	0.24	0.18
High-school Degree	0.26	0.25	0.22	0.29	0.23	0.22	0.31	0.20	0.25
Some College and Bachelors	0.26	0.30	0.26	0.27	0.31	0.32	0.24	0.34	0.23
Masters	0.03	0.02	0.04	0.06	0.09	0.13	0.08	0.14	0.18
Bus. Engin. School Degrees	0.02	0.06	0.05	0.03	0.11	0.08	0.08	0.08	0.16

Table 2.8: Probability of reaching an education level given the type and cohort: aggregation of education levels

	$p(h\|k,c)$								
Conditional on cohort ...	1998			2004			2010		
and conditional on type ...	1	2	3	1	2	3	1	2	3
High-school Degree and Less	0.69	0.62	0.65	0.63	0.49	0.46	0.60	0.44	0.43
Some College and Bachelors	0.26	0.30	0.26	0.27	0.31	0.32	0.24	0.34	0.23
Masters and School Degrees	0.05	0.08	0.09	0.09	0.20	0.22	0.16	0.22	0.34

is, $\hat{p}(h|k,c)$. This table confirms the observations already made, and also that the types are far from being completely characterized by their education level. As time passes, types seem to specialize more but all of them are characterized by shifts towards longer studies. To clarify the differences in educational 'choices' or (attainments) of the different types, Table 2.8 aggregates the degrees in three groups, with a clear hierarchy, and the essential phenomenon appears: 60% of type-1 students end up with a high-school degree or less in the last cohort, Bachelors (and the equivalent of Associates's degrees) are increasingly the common choice of type 2, while the type 3 (and to a lesser extent the type 2) are more concentrated at the top of the degree scale. It seems that the differentiation of type 2 and type 3 has increased with time (because their educational 'choice' patterns were very close in the 1998 cohort). Table 2.7 shows that there has been a rush on Master programmes and schools, and a certain flight from the lowest levels and the 'some college' category. The types differ only in the intensity of these changes.

The estimated parameters of the Multinomial Logit, describing education choices, are presented in Online Appendix B5, Table B.5. We discuss the impact of family background by type in Online Appendix B13.

2.6.4 Conclusion of the analysis of unobserved heterogeneity: who are the types 1, 2 and 3?

We can now summarize a reasonable interpretation of types.

1°) Type 1 has smaller returns to experience and smaller returns to education than other types. These individuals also tend to study less than other types. It seems that it is the group of individuals with a smaller ability.

2°) Type 2 occupies a median position in terms of returns to education, between Type 1 and Type 3, but closer to Type 3. Type 2 also occupies the median position in terms of returns to experience, but this time, closer to Type 1 than to Type 3. Type 2 is strongly characterized by a high employment rate, around 95%. To sum up, the Type 2 have a good level of ability and find stable jobs but their earnings grow slowly as compared to Type 3.

3°) Type 3 is clearly the 'top type' in the sense that these individuals are strongly characterized by markedly higher returns to experience. They also obtain the highest returns to degrees but have a much smaller employment rate than type 2, around 75%.

None of the types is easily predicted, or characterized by certain values of observable control variables, but we find some unsurprising correlations with family and geographical background variables.

2.6.5 Are unobservable types determined by, or correlated with, neglected observable characteristics?

By construction, types are supposed to be orthogonal to observable characteristics present in our estimation model. However, are they correlated with omitted pre-market variables? Can we observe the determinants of types?

To study this point, we used regularized regressions of type probabilities on pre-market variables, and more precisely, an *elastic net* method to select the variables correlated with type among all available controls. The elastic net is a *regularized regression* method that linearly combines the L_1 and L_2 penalties of the lasso and ridge regression methods. We first assign to each individual his most likely type (*i.e.*, the type with the highest ex-post probability). Then, using a ridge regression with a multinomial model where the hyper-parameters are estimated by cross-validation, we estimate which variables are correlated with the types. A first reassuring result is that observable variables do not help predicting correctly the types. The confusion matrix shows very poor prediction results. However,

among selected variables, results show that Type 2 and Type 3 are more prevalent among individuals who did not repeat a grade before junior high school, while Type 1 is more prevalent among individuals who have repeated a grade before high school. Type 3 is less prevalent among individuals who lived in rural areas at age 11 whereas types 1 and 2 are more prevalent among them. In addition, Types 2 and 3 are less prevalent among individuals whose parents work in agriculture, whereas Type 1 is more prevalent among members of this group. Type 1 is more common among individuals living in the south and west of France whereas Type 3 is more common among individuals living in Paris or the Paris region. Finally, Type 3 is also associated with individuals whose parents have a university degree. In Online Appendix B7, Table B.7 gives the results of the elastic net procedure. We conclude that types are somewhat correlated with some observable characteristics, but that types are not just dummies for omitted observable characteristics. For instance, family background is not a good predictor of types; the three types are present in all families.

2.7 Conclusion

In this article, we studied the evolution of wages during the early years of career of a large panel of individuals, in France. We stacked three surveys covering the first 7 years of career of young workers in France, from 1998 until 2017. The dataset takes the form of an unbalanced panel. We estimated a model describing the education choices, the accumulation of effective experience and individual wages simultaneously. Unobserved heterogeneity is handled by means of a finite set of latent individual types (a finite mixture model). Each type has its own Mincerian log-wage equation, its own employment-rate equation and education-choice model. The full model is estimated by means of standard Maximum Likelihood methods, using a sequential EM algorithm to find preliminary estimates. On top of a full set of type-dependent parameters, the estimation procedure yields the prior probabilities of types and, using Bayes' rule, the posterior probability that each individual belongs to any given type (*i.e.*, a probabilistic *classification* of individuals). From these results we compute policy-relevant parameters, such as the ATT and ATE of various education levels. The overall ATTs and ATEs can be expressed as averages of type-dependent treatment effects. So we obtain a representation of unobserved heterogeneity. This allowed us to show that the variation in time of the average real wages of workers, given a type of degree, is in some interesting cases

the average of devaluations for some types, and wage increases for other types. In a similar fashion, the returns to experience and experience accumulation are themselves heterogeneous. The devaluation (*i.e.*, absolute drops) of the real wages of Master's degrees holders is an average of divergent evolutions conditional on type. Overall, between 1998 and 2017, and after 7 years of career, the absolute variation of a Master degree holder's ATE is a drop of around 600 euros per month, if we treat the high-school dropouts as the untreated. The parallel variation of the ATT is a smaller drop of around 400 euros. We observe that the selection of students (or the quality of students) has improved with time in French Master programs, in spite of the growth in enrollment. We conclude that the observed devaluation is likely to be due to an excess supply of graduates because it cannot be attributed to a lesser average quality or productivity of the graduates.

Chapter 3

The Human Capital of Entrepreneurs: A Critical Review Using High Quality Danish Administrative Data

3.1 Abstract

In this review, we delve into the literature addressing the human capital of entrepreneurs. We draw comparisons between the insights derived from the previous literature and the high-quality information available in Denmark's administrative register. We first explore the various datasets that have been employed to investigate entrepreneurial human capital, outlining their respective strengths and limitations for comprehensively studying entrepreneurs. Subsequently, we shift our focus to theoretical frameworks concerning entrepreneurial human capital. We then delve into the empirical literature that has, over time, accumulated compelling evidence supporting the existence of such entrepreneurial human capital. Furthermore, we conduct a critical review of studies that have delved into the contents of entrepreneurial human capital. Throughout this analysis, we identify the areas where research has garnered significant traction and the gaps that present opportunities for future exploration. By juxtaposing the findings from previous literature with the information from Denmark's administrative register, we aim to provide a comprehensive and insightful overview of the evolving field of entrepreneurial human capital.

3.2 Introduction

This paper reviews the economics literature on the human capital of entrepreneurs. It initially presents Danish administrative data, which serve as a baseline for comparing the literature's findings with high-quality data. Subsequently, the paper addresses the matter of defining an entrepreneur in the data. There is no universally accepted definition of an entrepreneur, unlike the relatively clear distinction between a worker and other occupational roles. Furthermore, given the infrequent selection of entrepreneurship as a career path, researchers often face a dilemma between accurately defining entrepreneurship and the challenge of dealing with small sample sizes when identifying entrepreneurs in representative surveys. Moreover, the historical lack of data linkage between entrepreneurs and firm attributes has been a constraint. By meticulously reviewing both the data available and the definitions utilized for entrepreneurship within the literature, we scrutinize the available datasets for studying entrepreneurs, underscore their limitations, and propose a minimum criterion for designating a self-employed worker as an entrepreneur. Finally, considering the increasing availability of administrative data, we employ Danish administrative data to demonstrate how a substantial portion of registered incorporated businesses are not productive firms but rather legal entities. We elucidate the criteria essential for distinguishing genuine productive firms from those that merely function as legal constructs.

The fourth section of the paper is dedicated to reviewing theoretical models concerning the human capital of entrepreneurs. We commence with the Lucas model (Lucas (1978)), which examines how the distribution of managerial talent shapes firm sizes within an economy. Subsequently, we delve into its expansion, as presented in Guiso et al. (2021a), which discerns the distinction between entrepreneurial human capital necessary for business initiation and the one requisite for business establishment. We then examine the Lazear model (Lazear (2004a) and Lazear (2005)) concerning the multidimensional skill set of entrepreneurs, and we outline its implications for the composition of an entrepreneur's human capital vector. Lastly, we scrutinize the model introduced in Liang et al. (2018), which establishes a connection between the accumulation of entrepreneurial human capital and a country's demographic dynamics.

The fifth section of the paper addresses the evidence that has accumulated regarding the existence of skills relevant to entrepreneurship, which meet the criteria necessary to be classified as human capital. We first discuss the findings in Guiso

et al. (2021a)that provide compelling evidence of how individuals can acquire skills early in life, which later prove to be productive when engaging in entrepreneurship. Furthermore, we delve into the literature that attributes the inverted U-shaped pattern between age and the probability of becoming an entrepreneur to the accumulation of human capital relevant to entrepreneurship (Azoulay et al. (2020), Liang et al. (2018)).

Subsequently, the sixth section of the paper reviews the literature that has explored the components of entrepreneurial human capital. We delve into the empirical literature that attempts to provide evidence for the jack-of-all-trades theory (Lazear (2004a), Lazear (2005), Wagner (2003), Wagner (2006), Silva (2007)), while also emphasizing the limitations arising from data quality and the lack of causal evidence in establishing the relationship between the multidimensionality of human capital and the self-selection into entrepreneurship, as well as entrepreneurial outcomes. Additionally, we examine patterns of complementarity found in the relationship between certain skills and self-selection into entrepreneurship (Levine and Rubinstein (2017a), Michelacci and Schivardi (2020a)), highlighting how the search for patterns of complementarity in human capital can still yield promising findings. Finally, we discuss evidence regarding the relationship between the accumulation of entrepreneurial human capital and the traditional concept of managerial skills (Guiso et al. (2021a)).

The concluding section of the paper summarizes the outstanding inquiries surrounding entrepreneurial human capital and outlines prospective paths for future research endeavors within this domain.

3.3 The Data

In this section, we will describe how we constructed the dataset pertaining to firms and their owners by leveraging multiple administrative data sources from the Danish administrative register. The dataset employed is the same as the one utilized in Argan et al. (2023). For the sake of clarity, we present the dataset construction here, while the subsequent text mirrors that of Argan et al. (2023) exactly. The dataset combines multiple administrative data sources to formulate a comprehensive dataset encompassing all firm ownerships in Denmark spanning from 1996 to 2019. To achieve this, we combine individual-level characteristics from the Statistics Denmark Research Database (DST) with firm-level data from the Danish Central Business Register (CVR), along with commercially available data from

the KOB database (KOB) provided by Experian Denmark. This integration allows us to link individual level information to entrepreneurial spells and business outcomes.

The data contained in Statistics Denmark is provided and updated regularly by relevant Danish authorities, including the Ministry of Taxation, the Ministry of Education and the Ministry of Employment. The database contains general information on individuals such as gender, age, education, wealth and income composition. In addition, detailed employment registers provide all current and previous employment relationships (employer-employee), with corresponding salaries, hours worked, and occupational codes (isco 08) that are used to characterize individual labor market histories, as well as firm-level employment. However, the DST does not contain data on incorporated firms (limited liability companies), but only data on unincorporated firms (sole proprietorship and partnership). As shown by Levine and Rubinstein (2017a), when studying entrepreneurship it is key to separate between owners of sole proprietorships and owners of limited liability companies, as they display very different characteristics. To this end, we add the CVR database to the DST dataset, where the former contains information on all firms registered in Denmark since 1980. The CVR also contains detailed ownership records of sole proprietorships, partnerships and corporations and provides the timing, identity and ownership shares of all direct owners. As ownership records referring to incorporated businesses are limited to the period after 2014, we combine the CVR database with data from the commercially available KOB database, published by Experian Denmark, that contains hand-collected ownership information, which completes missing ownership in the early data years of the CVR database. The KOB database also contains detailed accounting records of corporations. All firms in the resulting dataset are identified by unique CVR-numbers, and all individuals are identified by unique PNR-numbers, which can be matched directly to other data sources.

After combining all these datasets we obtain a dataset in which the unit of observation is an individual. For every individual and annually for every year between 1996-2019 the final dataset contains information on individuals' income, net wealth, labor market status, hours worked, occupation, whether an individual owns a sole proprietorship, a partnership or a limited liability company and if so the corresponding business outcomes for each year in which the business exists: revenues, assets, number of employees, turnover, dividends and industry in which the business operates. Concerning years before 1996, the dataset contains, in addition to a variable on accumulated labor market experience, individuals' education

level (the highest educational attainment) and her type of education. In addition, for each individual the dataset reports demographic characteristics, the place of birth and the place of living. Finally, we are also able to link individuals with their parents and their siblings through the PNR number, if alive.

3.4 Identifying Entrepreneurs in the Data

Identifying entrepreneurs within the data presents a challenge due to the absence of a universally agreed-upon definition of who qualifies as an entrepreneur. However, various characteristics of individuals and their jobs have been utilized to discern entrepreneurs within the data, often due to the limited information available in datasets. The initial criterion for considering an individual an entrepreneur is their status as self-employed, as opposed to being an employee. This definition of self-employed individuals as entrepreneurs emerged early in the literature on entrepreneurial human capital, primarily due to data constraints. Given the scarcity of entrepreneur in the population, and thus in representative datasets and the aim to maximize observations, researchers opted for this broad self-employment categorization. Nonetheless, this loose definition inadvertently encompasses self-employed individuals without capital and employees who function as freelancers rather than entrepreneurs.A supplementary characteristic employed for identification is an individual's ownership, even partial, of a company—preferably an incorporated one. However, this criterion alone fails to distinguish between investors and individuals actively engaged in establishing or operating the business. A more precise criterion could involve an individual being the majority shareholder, thus possessing the authority to appoint management even if they aren't directly involved. An alternate scenario involves an individual being the founder of a company, indicating direct involvement in its establishment—an apt definition of an entrepreneur as an innovative contributor to the production process. Unfortunately, this information is rarely accessible in datasets. Furthermore, an individual who holds a C-level managerial position within a company might also be deemed an entrepreneur due to their entrepreneurial actions. Yet, when the individual is solely a manager and not an owner, the absence of personal capital investment, a key aspect of entrepreneurship, becomes evident. The optimal depiction of an entrepreneur encompasses ownership, founder status, top managerial role, self-employment, and an exclusive focus on the business. Nevertheless, this comprehensive definition is challenging to apply due to data limitations. Finally, even if she is the owner,

founder, and top manager of a firm, the firm is required to have assets, revenue, and employees and not be an empty legal shell. Considering the lack of consensus regarding entrepreneur definition and the associated data constraints, this discourse explores how recent empirical literature addresses the issue while acknowledging the impact of data availability on defining entrepreneurs.

Blanchflower and Oswald (1998) is one of the pioneering recent empirical studies in the field of entrepreneurship determinants . This study defines an entrepreneur as an individual who is self-employed and actively manages a business. The study articulates this definition with precision, expressing it as follows: the likelihood of engaging in business activities equals the product of the likelihood of possessing entrepreneurial vision and the combined likelihood of having access to capital or being capable of securing an unsecured loan in the absence of capital. As per the insights from Blanchflower and Oswald (1998), an entrepreneur is an individual who not only operates a business but also possesses the essential attributes of entrepreneurial vision and capital. However, upon transitioning to the empirical examination utilizing the National Child Development Study (NCDS), a longitudinal survey encompassing a cohort of individuals in Great Britain born between March 3 and 9, 1958, the definition of an entrepreneur takes a different form. In this context, an entrepreneur is classified as an individual labeled as self-employed—a categorization that might encompass a diverse range of individuals, potentially lacking both entrepreneurial vision and access to capital, not to mention the capacity to secure an unsecured loan. Both Lazear (2004a) and Lazear (2005) conceptualize entrepreneurs as individuals akin to ”jacks of all trades” who adeptly orchestrate essential production factors—human, physical, and informational—in an efficient manner. This definition portrays entrepreneurs as individuals who not only possess physical capital within a company, but also engage in workforce management, firm administration, and supplier interactions. Notably, this definition establishes a higher level of stringency compared to individuals with entrepreneurial vision and capital investment alone. It incorporates two additional criteria: the employment of workers and direct firm management. Delving into their empirical investigation, Lazear (2004a) and Lazear (2005) leverage a survey of Stanford MBA alumni to define entrepreneurs self-employed individuals owning an incorporated business. This definition stands as notably more stringent than the broader category of self-employed individuals. The incorporation distinction introduces several characteristics that harmonize with a more comprehensive definition of entrepreneurship, as elucidated in subsequent sections of this chapter. Wagner

(2006) follows Lazear (2004a) in defining an entrepreneur. However, utilizing data from the *Regional Entrepreneurship Monitor REM Germany 2003*, Wagner (2006) defines an entrepreneur as an individual who responds affirmatively to the question of whether she is actively engaged in initiating a new business that will be under their ownership, either in its entirety or partially. This particular definition of entrepreneurship emphasizes the concept that entrepreneurs inherently serve as the initiators and founders of novel ventures. Drawing from the *Longitudinal Survey of Italian Families; ILFI, 1997*, Silva (2007) categorizes self-employed individuals, regardless of whether they have employees, alongside managing partners of firms as entrepreneurs and omits family business owners. This expansive classification of entrepreneurs encompasses a wide spectrum of individuals, potentially pooling together those whose alignment with the core concept of entrepreneurship varies significantly. The category spans from self-employed individuals without any employees to managing partners of companies, potentially diverging from the fundamental concept of entrepreneurship.

The work conducted by Levine and Rubinstein (2017a) represents a significant breakthrough within the realm of entrepreneurship literature. The primary focus of this paper revolves around the task of defining an entrepreneur in the data. Through an examination employing data from the CPS (Current Population Survey) and NLSY79 (National Longitudinal Survey of Youth 1979), Levine and Rubinstein (2017a) delves into a comparative analysis of distinct categories of self-employed individuals: those who operate unincorporated businesses and those who oversee incorporated businesses. As can be seen in Figure 3.1 (Figure 1 of the paper Levine and Rubinstein (2017a)), individuals identified as owners of incorporated businesses (the rightmost column within Figure 3.1) exhibit positive selection concerning metrics such as earnings, working hours, and years of formal education, in comparison to salaried workers. Conversely, unincorporated business proprietors display contrasting trends: during their tenure as employees, they garnered lower earnings in contrast to traditional workers, devoted fewer hours to their work, and pursued fewer years of formal education. Importantly, these stylized patterns are consistently corroborated by both the CPS sample and the NLSY79 dataset. To provide further insight, unincorporated businesses tend to be characterized by a median firm size of zero and an average of 2.1 employees. Conversely, entrepreneurs who helm incorporated businesses oversee ventures that, at the median, encompass a staff of 2 employees, and on average 23 employees. In summation, the discernible trend is that while entrepreneurs steering incorporated businesses manifest positive

selection tendencies in terms of remuneration and educational attainment within the broader population, those at the helm of unincorporated businesses tend to be negatively selected along these dimensions. Moreover, it is noteworthy that unincorporated businesses predominantly represent individual undertakings, and their proprietors often lack the attributes associated with effective workforce management.

In Table 3.1, we present the average years of education and firm sizes pertaining to three distinct categories of business owners in Denmark from 1996 to 2019: sole proprietors, owners of partnerships, and proprietors of limited liability companies. Sole proprietorships can be likened to unincorporated businesses, partnerships represent unincorporated ventures designed for groups of professionals, and Limited Liability Companies denote incorporated enterprises. Through an analysis of Danish Administrative data, we have uncovered analogous trends to those highlighted by Levine and Rubinstein (2017a). Specifically, sole proprietors/unincorporated business owners exhibit markedly lower educational attainment in comparison to proprietors of limited liability companies, with partnership owners falling within an intermediate range. Turning our attention to firm sizes, we are able to reaffirm a consistent pattern: unincorporated businesses predominantly lack the attributes necessary for effective workforce management, in stark contrast to their incorporated counterparts.

Finally, Levine and Rubinstein (2017a) pushes the analysis further and show, using the U.S department of Labor's Dictionary of Occupational Titles, that activities performed by incorporated business demand strong non-routine cognitive skills that is creativity, problem solving, managing interpersonal relation, while unincorporated business are characterized by activities in the domain of manual skills. Furthermore, they demonstrate that it is not the case for a business to initiate as unincorporated and subsequently evolve into an incorporated entity. Hence, if our conception of an entrepreneur excludes, at the very least, professional self-employment or freelancing, then it becomes essential to omit unincorporated business owners from the entrepreneur's definition and concentrate on those incorporated businesses owners.

While Levine and Rubinstein (2017a) primarily focuses on the significance of incorporation as a means to distinguish non-entrepreneurial self-employed individuals, it neglects to address the inquiry of whether incorporated business owners serve as the actual founders of their firms and whether they actively manage these enterprises. In fact, Levine and Rubinstein (2017a) utilizes the CPS questionnaire,

Table 3.1: Education of owners and firm size of firms by legal status for the universe of Danish Company during the period 1996-2019. Number of Sole-Proprietorship business owner 377 107, Number of Partnership Business owner 53 281, Number of Limited Liability Company Business owner 102 439.

		Firm Size: Number of Employees		
	Education (Years)	Av.	Median	90 perc.
Sole-Proprietorship	12.27	0.18	0	0.4
	(.004)			
Partnership	12.89	1.4	0	0.3
	(.011)			
Limited Liability Company	13.08	3.9	0.2	8.7
	(.0078)			

which allows individuals to indicate if they are self-employed and whether their businesses are incorporated. However, the survey does not delve into the extent of their involvement in firm management or the specifics of their founding roles. As for the NLSY79 dataset, Levine and Rubinstein (2017a) leverages a survey question that inquires, "Do you consider yourself to be an entrepreneur?" Entrepreneur in this context is defined as someone who initiates a business endeavor, typically involving significant initiative and risk. While the NLSY79 approach presents an intriguing definition of entrepreneurship, the data collection hinges upon self-declaration, introducing limitations to the reliability of this measure. The study conducted by Azoulay et al. (2020) rigorously delves into the intricate delineation of an entrepreneur, with an emphasis on their role as both founder and manager. Employing the LBD (US Census Bureau's Longitudinal Business Database), the authors establish the notion of ownership by identifying individuals who hold partial or complete ownership of a firm at the inception stage (referred to as firm age zero). They meticulously disentangle the distinctions between investors and founders by cross-referencing ownership with active involvement in the firm's operations at its inception. In cases where specific firm ownership data is unavailable, Azoulay et al. (2020) defines entrepreneurs and founders by considering the three highest paid workers within the firm during its inaugural year. This work diligently crafts a definition of an entrepreneur that tightly aligns with the Schumpeterian concept of an entrepreneur as an innovator and driving force behind the firm's inception and management.Furthermore, Guiso et al. (2021a), through an analysis of the SHIW sample, a representative cross-section of Italian households, establishes an entrepreneur as an owner who actively oversees an incorporated business. In

the case of the ANIA sample, a survey encompassing over 2000 Italian firms with employee counts ranging from 10 to 250, the definition extends to encompass the individual in charge of firm management. These definitions of an entrepreneur encapsulate not only the legal framework of incorporation but also the managerial characteristics concerning employee management and overall business administration, as both definitions inherently require the individual at the helm of operational decision-making.

The Schumpeterian perspective on entrepreneurship asserts that entrepreneurs drive economic growth by replacing outdated enterprises with new and more efficient ones. These entrepreneurs undertake the risk of introducing novel goods, services, and production processes. Although this concept of entrepreneurship is not directly observable in data, economic literature over the past two decades has made significant progress in identifying such individuals within datasets. In this chapter, we explored how unincorporated businesses do not align with the Schumpeterian entrepreneurial definition, whereas incorporated businesses exhibit greater compatibility. Nonetheless, owners of incorporated businesses can also function as mere investors in a firm, lacking involvement in its founding or management. Thus, the literature underscores that individuals must not only invest risky capital in a venture but must also actively manage the firm by holding a top managerial position to be deemed entrepreneurs. To underscore the pivotal role of entrepreneurs in the growth process, the literature also distinguishes between those who acquire established ventures and true founders. The latter, often recognized as the individuals introducing new goods and processes, bear substantial importance. Researchers have long grappled with defining entrepreneurs due to data limitations. However, with the increased availability of administrative data, it has become feasible to identify individuals who both found and manage incorporated businesses. Nevertheless, a critical challenge lies in accurately discerning what constitutes a legitimate firm, as many registered incorporated entities serve as legal constructs without productive functions.

Utilizing Danish Administrative data, we present in Table 3.2 how nearly 50% of limited liability companies can hardly be considered as actual firms. Within this table, we showcase statistics for the universe of Danish limited liability companies, encompassing average firm duration and the number of individuals registered as owner based on three distinct attributes: hiring at least one employee during the observation period, recording positive revenue at least once, and registering positive assets at least once. Notably, only 143,235 out of 244,079 owners of lim-

ited liability companies in Denmark satisfy all three criteria during the observation period. In essence, when employing administrative data and adhering to the definition of an entrepreneur as an owner of an incorporated firm, approximately 42% of instances involve individuals who own businesses lacking employee engagement, revenue generation, or positive assets. While the most of dataset lacks data adequate for precise evaluation of asset and revenue positivity, Table 3.2 underscores that observing a firm hiring at least one employee serves as a fairly robust indicator of a genuine firm as opposed to a mere legal entity. Examining the bottom section of the table, specifically the last three rows designated with H(Y) (denoting hiring at least one employee during the observation period), it's evident that even though they might lack either revenue or assets, these instances encompass merely around 2,000 individuals—constituting less than 1% of the total limited liability company universe in Denmark. In contrast, the middle section of the table, characterized by rows marked with H(N) (indicating never hiring an employee during the observation period), accounts for nearly 40% of owners. Firms observed to never hire employees endure substantially shorter durations than actual productive firms, ranging from 2.3 years for those without employee engagement, revenue, or assets, to 2.5 years if only revenue is present, to 5.9 years for those with solely assets, and finally 6.5 years for firms exhibiting both assets and revenue. In conclusion, the pivotal takeaway is that the prerequisite of hiring at least one employee stands as a vital factor in identifying productive firms. This requirement furnishes a valuable and sufficient piece of information for discerning authentic productive enterprises.

Table 3.2: Average Duration of llc and number of individuals by categories

	Duration Years	Individuals
H(Y)R(Y)A(Y)	10.1	143 235
H(N)R(N)A(N)	2.3	18 961
H(N)R(Y)A(N)	2.5	3 715
H(N)R(N)A(Y)	5.9	18 505
H(N)R(Y)A(Y)	6.5	57 709
H(Y)R(N)A(N)	2.3	804
H(Y)R(Y)A(N)	2.6	329
H(Y)R(N)A(Y)	7.3	821
Mean	8.2	244 079

Entrepreneurs as limited liability company owners (244 079)
H: Hire at least one worker; H(Y) yes, H(N) no
R: Has at least once positive revenue; R(Y) yes, H(N) no
A: Has at least once positive asset; A(Y) yes, A(N) no
ex: H(Y)R(Y)A(Y) means H(Y) and R(Y) and A(Y)
Danish Administrative data 1996-2019

3.5 Theories of Entrepreneurial Ability

Lucas (1978) stands as one of the pioneering early works that addresses the notion that capital and labor do not seamlessly come together, highlighting the necessity of *managerial* ability within the production function.Within Lucas (1978), the author proposes the concept of a finite distribution of managerial ability within the population, which exhibits diminishing returns when applied to the production function. Due to diminishing returns to scale in managing both labor and capital, managers/entrepreneurs oversee a finite quantity of both capital and labor. Consequently, the distribution of managerial talent, given a specific set of production factors, molds the spectrum of firm sizes and establishes a threshold of managerial aptitude. This threshold dictates that individuals possessing lesser managerial talent remain employees, while those endowed with greater aptitude assume managerial roles.Given its primary focus on the theory of firm size distribution, the paper by Lucas (1978) does not extensively delve into the description of what constitutes managerial ability, nor does it expound upon how such ability is accumulated—it rather assumes a predetermined distribution. However, the paper tangentially addresses the observed phenomenon that individuals often transition into managerial roles later in their careers after a history of employment. Furthermore, the paper ensures that an extension encompassing accumulated managerial ability over an in-

dividual's employment trajectory does not influence the outcomes in terms of firm size distribution.

In contrast, Guiso et al. (2021a) propose an expansion of the Lucas (1978) model that precisely centers around managerial ability. They specifically differentiate between two aspects: the ability to reduce entry costs, which relates to entrepreneurial capacity essential for establishing a business, and the ability for skill improvement, necessary for efficient business operation. They contend that both these abilities can be acquired over time, a proposition empirically substantiated. In the context of career decisions, the two distinct abilities—business establishment and operational management—yield identical outcomes; namely, the higher these abilities, the more individuals are inclined to embark on entrepreneurial paths. However, when considering average performance, these abilities elicit contrasting effects. Indeed, while an increase in the ability to run a business within the population enhances business revenue and subsequently encourages more individuals into entrepreneurship, an elevation in the ability to establish a business reduces fixed entry costs. This reduction does not alter the inherent entrepreneurial ability, but rather allows individuals with lower entrepreneurial capacity to venture into entrepreneurship. Therefore, with an ascent in the population's ability to run a business, an elevation in the average quality of entrepreneurs becomes observable. Conversely, if the ability to establish a business improves, a decrease in the average entrepreneurial ability of entrepreneurs is observed. Nonetheless, in both scenarios, an increase in the number of entrepreneurs becomes evident. Guiso et al. (2021a) contribute by dissecting the managerial ability of the Lucas (1978) model into two components—establishing and running a business—thus offering testable implications.

In their study, Liang et al. (2018) construct a model aimed at exploring the relationship between a country's demographic makeup and its entrepreneurship rate. They propose a dual-pronged perspective. On one hand, the advantages of youth, driven by numerous factors such as dynamic social interactions and a propensity for innovative thinking that spawns groundbreaking ideas, that implies young individuals being better at entrepreneurship. Consequently, the study posits that entrepreneurship naturally declines as age advances. On the other hand, the model identifies another factor that elevates entrepreneurial prowess with age, which the authors term "business acumen". This expertise is acquired through hands-on experience within the workforce. A pertinent example involves on-the-job exposure within a particular sector, translating into valuable entrepreneurial experience within that domain. The complexity of this business acumen is further underscored by the need

for individuals to engage in intricate managerial tasks, rather than just any form of employment. By intertwining the benefits of youth with the concept of business acumen, Liang et al. (2018) successfully reproduce the characteristic inverted U-shaped pattern observed in empirical data, capturing the relationship between age and entrepreneurship rate. Moreover, by introducing the caveat that useful on-the-job experience for entrepreneurship is exclusively acquired at higher hierarchical levels within a firm, their model generates insights into the patterns seen in countries with higher median ages exhibiting lower rates of entrepreneurship .In expanding the characterization of managerial human capital, Liang et al. (2018) proposes that its acquisition hinges solely on the execution of intricate managerial tasks. Furthermore, they embrace the notion of a youthful advantage in creativity, portraying entrepreneurs as individuals brimming with disruptive, innovative ideas. Both Lazear (2004a) and Lazear (2005) delve deeply into the entrepreneurial human capital, elucidating its distinctions from the human capital possessed by workers. Their focus is centered on the premise that employees specialize in a singular task, while entrepreneurs encompass a multifaceted role involving activities like managing personnel, raising capital, and engaging with suppliers—effectively functioning as jack-of-all-trades, as per their definition. To formalize this notion, Lazear (2004a) and Lazear (2005) present the following framework: individuals possess a multidimensional skill vector, with each component representing distinct skills. Employees select the skill that generates the highest on-the-job revenue and earn corresponding wages based on this maximum skill's revenue. In contrast, entrepreneurs are compelled to leverage all skills within their vector, with their revenue constrained by the value of the least developed skill. This framework yields predictions in terms of self-selection into entrepreneurship. Firstly, individuals with a narrower discrepancy between their highest and lowest skill levels—termed a more balanced skill set—are more inclined to become entrepreneurs. This is rationalized by their decision-making process, wherein they compare the potential revenues as a worker, earning based on their top skill, against those as an entrepreneur, remunerated according to their weakest skill. Secondly, industries demanding a plethora of disparate skills are likely to harbor fewer entrepreneurs. This is attributed to the introduction of a correlation parameter across skills within an individual's skill vector. When skills are independent, the likelihood of possessing elevated levels of all such skills diminishes. Lastly, as skills are honed over an individual's job history, entrepreneurs have a distinct human capital accumulation strategy compared to those who remain workers. Specifically, entrepreneurs amass

expertise in areas where they lack proficiency, fostering a diversified human capital accumulation strategy unlike that of workers.
In this section, we have examined the primary theories presented in the literature. In the subsequent sections, we will delve into the empirical evidence concerning human capital of entrepreneurs.

3.6 The Empirical Evidences on Entrepreneurial Human Capital

There are three requirements to define a skill as human capital: 1) it must be productive, 2) it should be costly to develop, and 3) it needs to be retained within the individual. In this section, we will review the empirical literature that highlights the existence of entrepreneurial skills meeting these criteria. These entrepreneurial skills enhance individuals' productivity as entrepreneurs while satisfying the aforementioned three conditions.
The study by Guiso et al. (2021a) specifically addresses the question of whether acquirable skills exist for entrepreneurship. Leveraging Italian data on entrepreneurs and their firms (data are discussed in the section identifying entrepreneurs in the data), the authors reveal that individuals raised in areas with a higher concentration of firms possess an higher likelihood of becoming entrepreneurs later in life. To tackle the endogeneity concern arising from the fact that areas with higher entrepreneurial density tend to persist over time and might induce entrepreneurship for various unrelated reasons, they examine the subsample of movers. These are individuals who grew up in regions with high entrepreneurial density but moved before starting a firm. By utilizing this mover subset, the authors break the correlation between past and present firm densities. The authors conclude that the individuals who move bring with them a set of skills that substantially enhance their probability of becoming entrepreneurs. Furthermore, Guiso et al. (2021a) shows that sectorial distribution during the learning age determines the choice of a specific sector for entrepreneurship. This holds significance due to the sector-specific nature of certain skills, which serves as evidence supporting the transmission of entrepreneurial skills. As previously discussed in section 3 (theories of entrepreneurial human capital), Guiso et al. (2021a) introduces a model encompassing two distinct types of entrepreneurial human capital: one that is geared toward business establishment and another tailored for effective business operation. Utilizing data on firm performance, they illustrate that a higher firm density during the learning phase

enhances overall firm performance. This leads them to the conclusion that the interaction with entrepreneurs and firms during one's early years primarily nurtures the capability to effectively manage a business, rather than merely establishing it. In essence, Guiso et al. (2021a) provides compelling evidence substantiating the presence of entrepreneurial human capital, drawing insights from data encompassing Italian entrepreneurs and firms.

A notable observation that pertains to the accumulation of human capital in entrepreneurial capability is the connection between the likelihood of becoming an entrepreneur and age. In fact, the proportion of individuals engaged in entrepreneurial activities increases with age, and the relationship between probability to become an entrepreneur and age takes the shape of an inverted U. This stylized phenomenon is illustrated in Figures 2 and 3, where we present this trend using administrative data from the Danish population. This phenomenon has also been documented in the study by Azoulay et al. (2020), and in Liang et al. (2018). The upward trajectory of entrepreneurship with age, as discussed in Section 3, has been attributed by Liang et al. (2018) to two opposing forces: the advantage of youth and the accumulation of entrepreneurial skills over time. However, while Liang et al. (2018) present evidence consistent with the predictions of their model, but they are unable to directly test the underlying mechanisms governing the relationship between age and the likelihood of becoming an entrepreneur. On the other hand, Azoulay et al. (2020) offer compelling evidence supporting the notion of accumulating entrepreneurial human capital with age. They demonstrate not only an increase in the probability of individuals becoming entrepreneurs with age, but also a substantial rise in the performance of startups led by older founders. Figure 3.4 reproduces Figure 2 from Azoulay et al. (2020). Panel A showcases the likelihood of achieving a successful exit by the founder's age. The probability of a successful IPO for firms founded by individuals over the age of 50 is three times greater than that for firms founded by individuals in their twenties. A similar trend emerges when examining the likelihood of a firm reaching the top 0.1% employment rate within 5 years of establishment. While the increasing likelihood of entering entrepreneurship as age advances is in line with the concept of liquidity constraints – suggesting that individuals need time to accumulate savings and secure loans – the observed outcomes are challenging to attribute solely to liquidity constraints. This is due to the fact that individuals have already initiated their ventures and presumably possess the financial means to operate them. While personal wealth might aid in the successful operation of a firm, the pronounced rise in firm performance with age is difficult

to explain solely through this lens. It strongly indicates that individuals who have accumulated substantial human capital through professional experiences are better equipped, from a human capital perspective, to effectively manage and operate a business.

As we discuss the evidence regarding the existence of entrepreneurial human capital, it becomes evident that the valuation of such human capital can fluctuate over time. Michelacci and Schivardi (2020a) examines how returns from entrepreneurship vary over time based on different levels of owner education. In Figure 3.5 (Figure 2 of Azoulay et al. (2020)), they illustrate how returns from entrepreneurship for individuals with post-graduate degrees have experienced significant growth over the period 1995 to 2015. The paper addresses the question of potential compositional changes that may have occurred among post-graduate individuals and demonstrates that the results remain robust even when accounting for variations in the composition of the post-graduate population over time. This evidence suggests that education-related human capital holds substantial value for entrepreneurship, and over the past 30 years, this value has increased, particularly among individuals with higher levels of education.

In this section, we have examined the evidence indicating the presence of entrepreneurial human capital. In the following section, we will explore the evidence that delves into the composition of this entrepreneurial human capital.

3.7 What's Entrepreneurial Human Capital?

In this section, we will review and discuss empirical research that has aimed to unpack the concept of entrepreneurial human capital. The pioneering works of Lazear (2004a) and Lazear (2005) point out that entrepreneurial human capital does not arise from a specific set of skills. Instead, it concerns the characteristics of an individual's human capital vector. Specifically, it involves the minimum amount of skills required for the productive process that define the boundaries of entrepreneurial human capital. In the study by Lazear (2004a), Lazear leverages education records of Stanford MBA alumni in conjunction with their employment history (detailed in the data section). The study demonstrates that MBA alumni with more specialized educational backgrounds—where the number of courses taken outside their primary field is limited—have a lower likelihood of pursuing entrepreneurship later in life.While the evidence is suggestive, the degree of specialization is operationalized based on the count of courses taken out-

side one's field of expertise. This variable lacks clear economic significance and presents challenges in interpreting coefficient values. Furthermore, the sample of Stanford MBA students is highly specific and lacks external validity due to its non-representativeness in dimensions likely relevant to the decision to engage in entrepreneurship. In Lazear (2005), Lazear contributes further evidence regarding the "jack-of-all-trades" phenomenon by continuing to examine Stanford MBA alumni. This time, he utilizes the panel data of their job histories, using the number of roles they have occupied over their careers as a measure of specialization. These roles are self-declared by the individuals. The data indeed indicates that individuals who have held multiple roles throughout their job histories are more likely to become entrepreneurs later in life. Despite the aforementioned dataset limitations, this measure of specialization represents an advancement compared to the earlier approach involving a mix of chosen courses. However, it still retains the potential for diverse interpretations. Individuals who have held numerous roles might also include those who have rapidly advanced in their careers, achieving high positions and responsibilities at a relatively young age. It's evident that another limitation of this outcome lies in its observational nature, and the issue of endogeneity remains unaddressed.

In Wagner (2003), the investigation of the "jack of all trades" theory takes a step further by utilizing a sample that is representative of the German population, thereby representing an improvement over the sample used in Lazear (2004a) and Lazear (2005). Nevertheless, the variable employed to quantify multidimensionality remains tentative. Specifically, it employs the number of professional training programs completed after schooling, along with the number of times an individual has changed professions. Both of these measures—completion of professional training post-schooling and frequency of changing professions—positively correlate with the probability of self-employment in the data. However, the findings are constrained due to the arguable nature of the proxy for multidimensionality, and the failure to address the endogeneity issue.In Wagner (2006), using a distinct dataset also gathered in Germany, the researchers try to refine the proxy for multidimensionality. They achieve this by capitalizing on a question that directly asks respondents about the various professional fields they have been engaged in previously, clarifying that this inquiry pertains to the number of diverse fields and not the number of employers they have worked for. This variable, indicating the number of fields of experience, displays a positive correlation with the likelihood of an individual being a nascent entrepreneur in the data—someone in the process of

establishing a business. Although this variable to capture multidimensionality represents an advancement at the time compared to existing literature, it still remains unsatisfactory. The issue of endogeneity is left unattended, and the empirical analysis suffers from a limited sample size of entrepreneurs, encompassing only 168 nascent entrepreneurs within a broader sample of approximately 5000 individuals.
The study by Silva (2007) aims to confront the limitations posed by observational evidence within the empirical literature on the multidimensional human capital of entrepreneurs. To address these limitations, the author leverages the longitudinal dimension of the ILFI dataset (refer to the data section) and employs a fixed-effect identification strategy. As a proxy for multidimensionality of skill sets, the study adopts the concept of roles assumed over an individual's career. To define an entrepreneur, the study considers self-employed individuals, with or without employees. The findings of Silva (2007) demonstrate a positive correlation between varying numbers of roles and the likelihood of self-employment when employing pooled OLS regression. However, this correlation vanishes once the analysis accounts for endogeneity using the fixed-effect approach. Consequently, Silva (2007) concludes that these outcomes suggest inherent "jack-of-all-trades" tendencies in individuals either driven by a preference for diversity or innate skills that enable them to assume diverse roles in life and rendering them more inclined towards entrepreneurship. While Silva (2007) is the pioneering study in addressing the endogeneity issue within the "jack-of-all-trades" literature, the evidence remains constrained on multiple fronts. On one hand, the utilization of fixed effects as an identification strategy assumes that self-selection can be modelled by a constant factor. On the other hand, the definition of an entrepreneur closely resembles that of a self-employed individual. Lastly, the measurement of multidimensionality through the number of roles undertaken throughout life presents limitations.
The research by Levine and Rubinstein (2017a) delves into the characteristics of incorporated business owners and their businesses, offering insights that shed light on the composition of entrepreneurial human capital. Their study reveals that incorporated business owners, as well as their businesses, engage in tasks requiring higher levels of nonroutine cognitive abilities compared to both unincorporated business owners and workers. This task-job alignment is determined using the U.S. Department of Labor's Dictionary of Occupational Titles, a widely employed tool in economics literature to correlate skills with job roles. By employing data from both the CPS and NLSY79, they demonstrate that incorporated business owners perform better on learning aptitude tests at a young age, exhibit above-average

self-esteem and a sense of control over their futures, and tend to engage in a higher frequency of illicit activities. Interestingly, they also find that the most influential predictors for becoming an incorporated business owner in their dataset is the complementarity between high scores in learning aptitude tests and prior participation in illicit activities during youth. While some of these traits, as discussed above, might not be considered conventional human capital since they are acquired without deliberate effort, Levine and Rubinstein (2017a) suggests a complementarity between cognitive skills and a rule-breaking attitude. However, this interpretation somewhat stretches the concept of illicit activities and remains rooted in observational findings, lacking the resolution of the endogeneity concern. Nevertheless, it's noteworthy that this pattern of complementarity between skills or traits is a consistent feature in the self-selection process of entrepreneurship. On a more conventional definition of human capital, Michelacci and Schivardi (2020a) investigates how the complementarity between education and job market experience affects positevely returns from entrepreneurship. They interpret education as theoretical competencies and labor market experience as practical expertise gained through work experience. Utilizing the SCF (Survey of Consumer Finance), they reveal a strong interdependence between education and job market experience in explaining returns from entrepreneurship, with this interdependence growing markedly over time, particularly when comparing the period before 2000 to that after 2000.

Evidence highlighting the significance of skill complementarity in shaping self-selection into entrepreneurship has been amassing within the literature. These studies explore various skill combinations and their pivotal role in influencing the decision to engage in entrepreneurship. However, research focusing on specific skills aligned with the classical notion of effective business management remains comparatively scarce. An exception is Guiso et al. (2021a), which endeavors to unpack the constituents embedded in the entrepreneurial human capital developed during youth. The study reveals a noteworthy positive correlation between entrepreneurial exposure during formative years—rather than current entrepreneurial exposure—and several measures of managerial practices. In particular, the study identifies a positive link between entrepreneurial exposure during the learning age and measured managerial practices encompassing monitoring (the ability to oversee performance and assess outcomes), target setting (the capacity to define both qualitative and quantitative objectives), and people management (pertaining to human resource supervision). Thus, the research by Guiso et al. (2021a) offers compelling evidence supporting the existence of entrepreneurial human capital that sig-

nificantly influences an individual's aptitude for managing a business. Importantly, this human capital appears to be learnable. The study contributes observational evidence indicating that such human capital entails the ability to effectively manage individuals, set targets, and monitor and review performance.While Guiso et al. (2021a) presents a crucial contribution as the sole economic study to date on this subject.

3.8 Conclusion

In this paper, we provide a comprehensive review of the literature concerning the human capital of entrepreneurs. We begin by examining what nowadays are the minimum requisites for accurately defining an entrepreneur, which now encompasses owners of incorporated businesses employing at least one individual. Subsequently, we delve into an exploration of theoretical perspectives on entrepreneurial human capital, tracing the progression from the foundational work of Lucas (Lucas (1978)), which posits a distribution of managerial talent within the population. This leads us to more recent models of entrepreneurial human capital, including the differentiation between skills needed for business establishment and those necessary for operational management (Guiso et al. (2021a)). We also analyze Lazear's model on the multidimensional human capital of entrepreneur (Lazear (2004a)), and Liang's model (Liang et al. (2018)), which emphasizes the combination of youthful advantages with the accumulation of business acumen through on-the-job managerial experience. We then highlight the burgeoning body of evidence that has coalesced to support the existence of entrepreneurial human capital.
In conclusion, the paper discusses the empirical literature concerning the substance of entrepreneurial human capital. Notably, we address the current lack of robust empirical evidence supporting the Lazar's Jack-of-all-trades theory. Presently, there is a paucity of studies where the identification of entrepreneurs within the data is fully satisfactory, and the variables employed to proxy the multidimensionality of human capital are notably imprecise. Furthermore, no research has convincingly addressed the causal impact of heightened multidimensional skill sets on the likelihood of entering entrepreneurship. Moreover, no investigation has examined whether individuals with greater multidimensionality are not only more likely to become entrepreneurs but also more productive in their entrepreneurial endeavors. Hence, generating compelling empirical evidence related to the Jack-of-all-trades theory presents a promising avenue for further research. Addition-

ally, we delve into observed patterns of complementarity between different skills documented in the literature. For instance, there is evidence of a synergy between cognitive skills and a disposition for illicit activities (Levine and Rubinstein (2017a)), as well as a link between education and practical experience (Michelacci and Schivardi (2020a)). While these findings are compelling, it is logical to consider that other forms of complementarity are pertinent to self-selection into entrepreneurship and the productivity of entrepreneurs. For instance, one could hypothesize a complementary relationship between financial skills and human resource management. It's also plausible that the array of skills required for effective business management might be acquired through collaboration among shareholders. Notably, empirical research investigating these aspects is absent in the field of economics, suggesting that exploration of these topics could yield valuable research insights. Lastly, we address the empirical evidence presented by Guiso et al. (2021a) concerning human capital related to classical managerial practices. Nevertheless, this remains the sole piece of evidence on this subject. The investigation into whether entrepreneurs inherently possess management skills or acquire them over the course of their careers, and the means by which they do so, presents another avenue open for exploration within the realm of research.

Figure 3.1: Picture of Table 1 pag 970 from the paper Smart and Illicit: who becomes entrepreneur and do they eran more by Ross Levine and Yona Rubinstein

DEMOGRAPHICS AND LABOR MARKET OUTCOMES BY EMPLOYMENT TYPE

			Self-employed		
	All	Salaried	All	Unincorporated	Incorporated
Panel A: CPS 1996–2012					
Observations	1,225,886	1,108,591	117,295	75,476	41,819
	100.0%	90.4%	9.6%	6.2%	3.4%
A. Labor market outcomes					
Mean earnings	$ 47,515	$ 46,421	$ 58,174	$ 40,820	$ 89,169
Median earnings	$ 36,090	$ 36,363	$ 34,190	$ 24,625	$ 55,591
Median hourly earnings	$ 18.0	$ 18.0	$ 17.4	$ 13.8	$ 24.6
Annual hours worked	1,985	1,976	2,078	1,936	2,331
Full-time, full-year	0.69	0.70	0.64	0.57	0.78
B. Demographics					
Age	40.2	40.0	42.9	42.4	43.6
White	0.70	0.69	0.79	0.76	0.83
Female	0.48	0.49	0.36	0.40	0.28
Years of schooling	13.7	13.7	13.9	13.6	14.5
College graduate (or more)	0.33	0.33	0.36	0.31	0.46
Panel B: NLSY79 1982–2012					
Observations	132,681	121,782	10,899	8,963	1,936
	100.0%	91.8%	8.2%	6.8%	1.5%
A. Labor market outcomes					
Mean earnings	$ 44,725	$ 43,605	$ 55,785	$ 45,713	$ 93,411
Median earnings	$ 35,170	$ 35,222	$ 33,965	$ 28,672	$ 61,424
Median hourly earnings	$ 17.2	$ 17.2	$ 16.8	$ 14.7	$ 26.2
Annual hours worked	1,966	1,953	2,088	1,991	2,461
Full-time, full-year	0.59	0.59	0.53	0.48	0.72
B. Demographics					
Age	36.2	36.0	38.1	37.5	40.1
White	0.81	0.80	0.87	0.86	0.90
Female	0.47	0.48	0.38	0.41	0.28
Years of schooling	13.8	13.8	13.6	13.4	14.2
College graduate (or more)	0.30	0.30	0.26	0.23	0.36
C. Firm size: number of employees					
Median			0.0	0.0	2.0
Mean			8.6	2.1	23.0

Notes. The table presents summary statistics from the March Annual Demographic Survey files of the Census Bureau's CPS for the work years 1995 through 2012, for prime age workers (25 through 55 years old), and from the Bureau of Labor Statistics' National Longitudinal Survey of Youth 1979 (NLSY79) for workers who are least 25 years old between 1982 and 2012. The CPS and the NLSY79 classify workers in each year as either salaried or self-employed, and among the self-employed, they indicate whether the person is incorporated or unincorporated self-employed. The number of employees includes all paid employees in the year that the person becomes full-time self-employed and excludes the self-employed business owner, which is available from 2002 onward in the NLSY79. When using the CPS, we further exclude observations with missing data on age, race, gender, schooling, industry codes, or occupation codes, and those living in group quarters or working in agriculture or the military. When using the NLSY79, we further exclude observations with missing values on age, race, or cognitive and noncognitive traits (AFQT, Rosenberg Self-Esteem and Rotter Locus of Control). The Online Data Appendix provides further details on the sample and variables.

Figure 3.2: Share of Entrepreneur by age in cohort in three year of birth cohort of Male Danish Citizen: Cohort 1962,1968, 1974. Data from Danish Administrative register-see data section

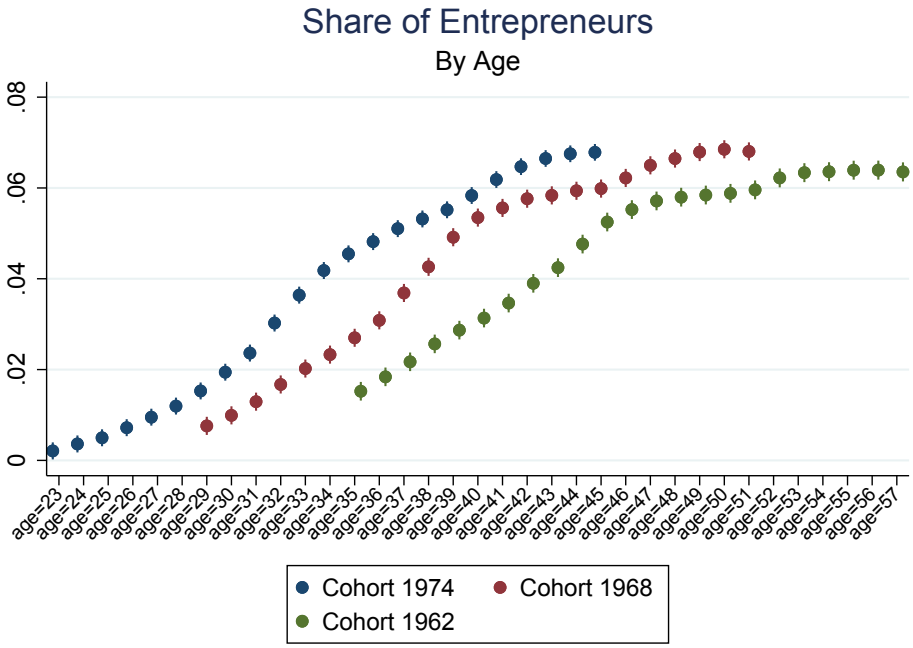


Figure 3.3: Share of Individual becoming entrepreneur by age in the population of Danish male born between 1961 and 1975

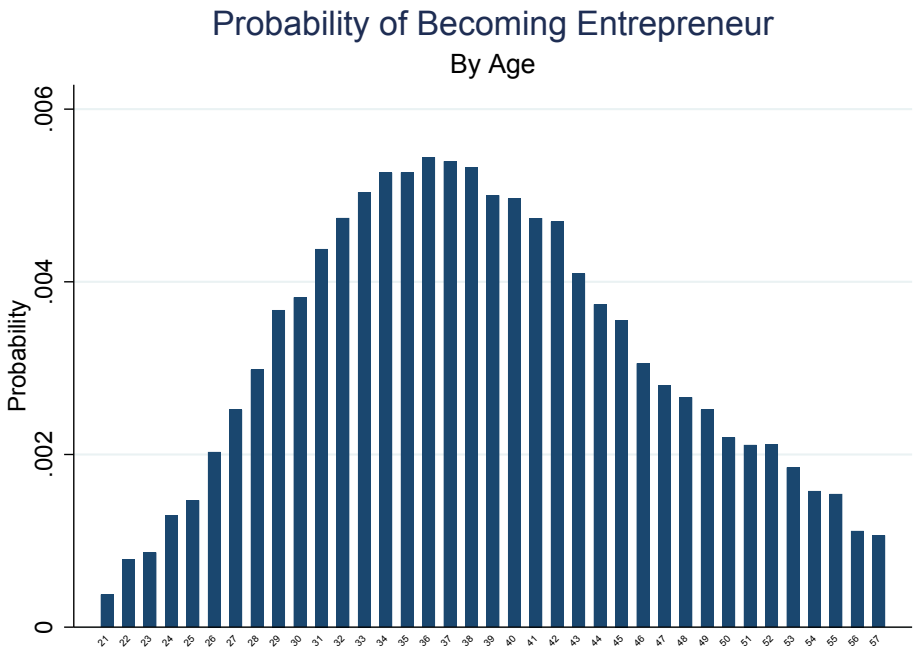


Figure 3.4: Figure 2 pag. 74 from the paper Age and High-Growth Entrepreneurship by Azoulay et al. , AER: Insight 2020

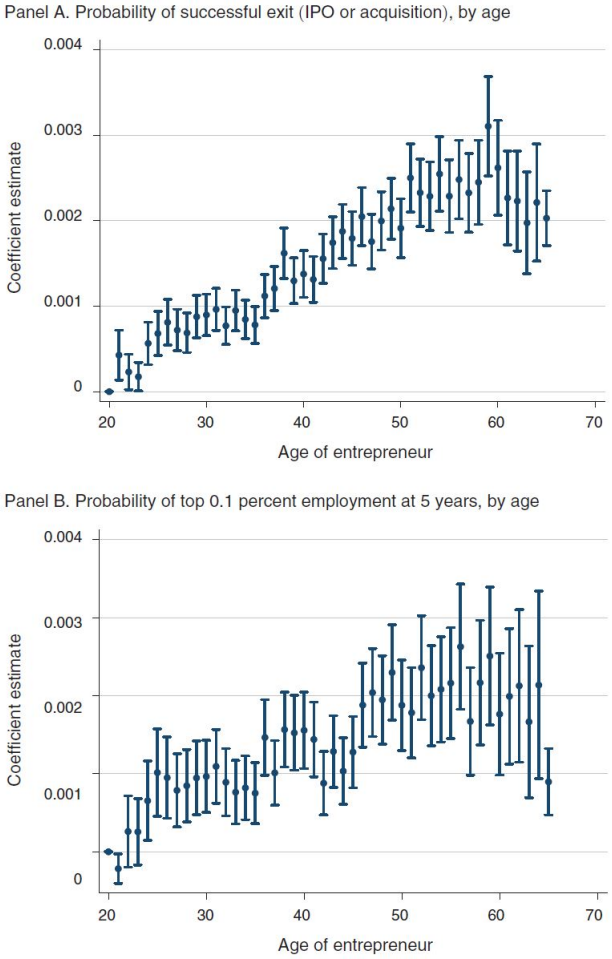


FIGURE 2. LIKELIHOOD OF EXTREME SUCCESS, CONDITIONAL ON STARTING A FIRM

Notes: OLS regression coefficients from estimating the likelihood of extreme firm success on a series of age indicators are shown. Ages 20 and below are grouped as 20 while ages 65 and above are grouped as 65. IPO data are sourced from Compustat. Acquisitions are based on firm ownership changes in the LBD. Top 0.1 percent employment outcomes are calculated based on five-year employment growth in the LBD. Regressions use robust standard errors.

Source: Authors' calculations based on W-2 earnings records, form K-1, LBD, and Compustat for firms founded over the 2007–2009 period.

Figure 3.5: Figure 2 from paper Are they all like Bill, Mark, and Steve? The education premium for entrepreneurs by Michelacci et al. 2020, Labour Economics

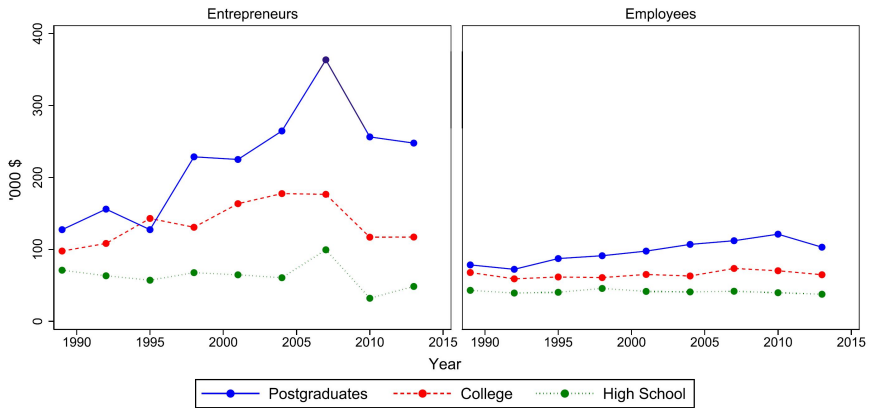


Fig. 2. Entrepreneurs' returns θ and employees' wage income w.
Source: Own calculations using the Survey of Consumer Finances, the Longitudinal Business Database and the S&P500 Total Return Index. Values are in thousands of dollars at constant 2010 prices.

Chapter 4

Teach the Nerds to Make a Pitch: Multidimensional Skills and Selection into Entrepreneurship

Co-written with **Leonardo Indraccolo** and **Jacek Piosik**

4.1 Abstract

Using Danish administrative data, in this paper we study how individuals' skill set composition affects self-selection into entrepreneurship. We use detailed education registry data on high school grades to measure individuals' human capital. Specifically, we measure analytical and communication skills with high school grades in math and the one in Danish language. We find that the final year high school GPA, the final average grade in math and Danish are negatively associated with the probability of becoming an entrepreneur. However, we observe a positive complementarity between math and Danish language grades in predicting individuals' self selection into entrepreneurship. In particular, for students with high math grades the probability of starting a business is monotonically increasing in their oral grade in Danish, while it is not so for the rest of the population. We interpret these observational findings as evidence of the importance of a balanced skill set, particularly in the complementarity between analytical and communication skills, for self-selection into entrepreneurship. For the population of high performing math students, we propose an identification strategy to casually estimate the effect of increasing communication skills on the probability of becoming an entrepreneur. We use information on parents' human capital and exploit within-school, across-

cohort variation in students' exposure to peers whose father has a university degree in humanities. We find that the most treated individual (90th percentile) in our sample has 1.1 percentage points higher probability of becoming an entrepreneur compared to the least treated one (10th percentile). The effect is economically significant, being equal to 20% of the overall share of entrepreneurs in the economy. Motivated by the evidence that students performing well in math run on average more profitable and bigger businesses, we highlight the importance of improving communication skills of individuals with high analytical abilities to incentivize the creation of high performing firms.

4.2 Introduction

Which skills do individuals selecting into entrepreneurship have and how does the skill set composition of entrepreneurs differ from the one of workers? Other studies, starting from the seminal contribution by Lazear (2004b), have investigated the role of human capital and skill set compositions as determinants of self selection into entrepreneurship. However, the question remains largely unanswered because of the unavailability of detailed datasets on entrepreneurs and their related characteristics. [1]

Our contribution is to use danish education registry data, combined with other administrative data sources, to quantify and casually estimate the effect of the complementarity between analytical and communication skills on the probability of self-selecting into entrepreneurship.

We have access to a rich and detailed dataset that combines multiple administrative data sources to generate a unique dataset that contains all firm ownership in Denmark between 1996 and 2019. Our dataset allows us to follow individuals over their life-cycle and observe their characteristics before they start a business, during their entrepreneurial spell and after. We also observe the same information for paid employed workers, which enables us to study differences in skill set compositions between workers and future entrepreneurs. Motivated by the theory of Lazear (2004b), we use high school grades in math and danish language in the last year of high school to measure communication and analytical skills. We start by providing observational evidence on the complementarity vs substitutability of these skills on

[1] Queiró (2022) uses Portuguese administrative data to show that more educated entrepreneurs run bigger businesses which display higher growth rates at the beginning of their life-cycle. Michelacci and Schivardi (2020b) use data from the SCF to calculate the returns to education for entrepreneurs.

the labor market and for selection into entrepreneurship. We find that individuals with more specialized skill sets earn higher wages, but are less likely to start a business compared to individuals with more balanced skill sets. We also show that for the group of individuals who were very good in math during high school the probability of becoming entrepreneurs is monotonically increasing in their oral grade in danish, while the same does not apply for the rest of the population of students. Additionally, we find that entrepreneurs who performed well in math during high school run on average more profitable and successful firms.

Understanding how policymakers can incentivize the creation of new successful businesses motivates us to casually estimate the effect of improving communication skills of students with high mathematical grades in high school. Being able to precisely quantify the increase in the number of new businesses created by individuals with high analytical abilities when these are taught better communication skills, is crucial when designing effective training and education programs aimed at spurring the creation of high performing firms.

Our identification strategy draws from the literature on peer effects and uses information on parents' human capital. We exploit within school, across cohort variation in the share of schoolmates parents' with an academic background in humanities. Motivated by the fact that human capital and skills get transmitted across generations, we instrument communication skills of high performing math students with the human capital composition of parents' peers students. For the cohorts of students graduated between 1997 and 2004, the estimated effect of increasing the share of fathers' peers with a university diploma in humanities by 3.5% - corresponding to the difference between the most and the least treated individual in our sample- increases the probability of selecting into entrepreneurship by 1.1 percentage points. The effect is economically significant if one considers that the share of individuals who ever become entrepreneurs in the economy for the 1997-2004 cohort is 4.8%. The effect corresponds to an increase of 20% in the overall share of entrepreneurs.

Other findings by Guiso et al. (2021b) have highlighted the importance of exposing aspiring entrepreneurs to entrepreneurial environments to stimulate the birth of new successful ventures. To put our results into perspective, we show that the estimated effect of improving communication skills of highly talented math students on the creation of new businesses is comparable in magnitude to increasing students exposure to a higher share of peer students whose fathers are graduated in business.

The rest of the paper is organized as follows. The next section discusses our con-

tribution in relation to other work on entrepreneurship. The third section describes our dataset in detail. The fourth and fifth section provide observational evidence on the role of gpa and high school grades on labor market outcomes and selection into entrepreneurship. The fifth section provides evidence on the complementarity between communication and analytical skills in self-selection into entrepreneurship. The sixth section discusses how high ability math students run on average more profitable businesses. Section seven introduces the identification strategy, section eigth discusses the results and the final section concludes.

4.3 Related Literature

This paper contributes to different strands of the literature. We contribute to the empirical literature on the role of human capital and skills for the understanding of selection into entrepreneurship (Lazear (2004b), Lazear (2005), Wagner (2003), Wagner (2006), Silva (2007),Levine and Rubinstein (2017b), Michelacci and Schivardi (2020b)) and to empirical studies investigating whether skills relevant to entrepreneurship are learnable (Guiso et al. (2021b), Liang et al. (2018)).
Prior work has been concerned with bringing empirical evidence to the model of entrepreneurship developed by Lazear (2004b). The model predicts that individuals selecting into entrepreneurship must have a more balanced skill set compared to workers (entrepreneurs are jack-of-all traits individuals). The result stems from the assumption that entrepreneurs are required to perform very different task using a variety of skills, while workers only need to perform very specialized tasks. Lazear (2004b) tests the theory empirically exploiting the Stanford MBA alumni register. He shows that students with less specialized course tracks, namely individuals who took more courses outside of their track of specialization, are more likely to become entrepreneurs in the future. Additionally, Lazear (2005) uses the same dataset to show how MBA alumni who had a higher number of different occupations as workers are more likely to become entrepreneurs later in life. Along similar lines, Wagner (2003) uses a representative sample of the German population and shows that individuals with more professional training in life, or who changed their profession more often, are more likely to become self-employed later in life. Wagner (2006) uses German data on nascent entrepreneurs (*Regional Enrepreneurship Monitor REM Germany*) to show that individuals who reported to have been active in many different professional fields are more likely to be observed as nascent entrepreneurs. Finally Silva (2007), using the Longitudinal Survey of Italian Families

(ILFI, 1997) and exploiting the panel dimension of the data, shows that while the total number of occupations an individual had in his career is positively associated with the probability of being self-employment later in life, the relationship disappears once accounting for the endogeneity by means of fixed effects .

Our research contributes to this literature along several dimensions. First, by using Danish administrative data we define an entrepreneur as the owner of an incorporated business that is economically active. Moreover, we work with the universe of Danish entrepreneurs. This represents an improvement in the quality of the data used to study the determinants of entrepreneurship. Prior work, as Lazear (2004b) and Lazear (2005), used survey data not representative of the general population. On the other side, Wagner (2003) does not distinguish between entrepreneurship and self-employment, while later work by Levine and Rubinstein (2017b)) has shown how this is critical, as the two groups have very different characteristics. Second, using education registries we have a clear and interpretable measure of skills, namely grades in math for analytical skills and grade in oral Danish examinations for communication skills. Previous literature used very imprecise proxies for the measurement of skills, such as the number of courses taken outside the field of specialization (Lazear (2004b)) or the number of different prior occupations (Lazear (2005), or the number of professional degrees (Wagner (2003)). Finally, we address the problem of endogeneity in the relationship between multidimensional skill sets and the decision to become an entrepreneur, which so far has only been addressed by Silva (2007) using panel fixed effects.

The literature studying the human capital determinants of entrepreneurship and firm performance has found patterns of complementarity between skills affecting both self selection into entrepreneurship and firms outcomes. In particular, Levine and Rubinstein (2017b) show how having high levels of cognitive ability coupled with the tendency to perform illicit activites when young is a very strong predictor for selection into entrepreneurship. The authors interpret this finding as the existence of complementarity between cognitive skills and the tendency to break established rules. Michelacci and Schivardi (2020b) show that the education-labor experience complementarity is associated with higher returns from entrepreneurship and interpret this finding as the existence of complementarity between theoretical knowledge acquired during formal studies and practical skills acquired on the job. We add to these findings by showing evidence of complementarity between analytical and communication skills.

Finally, we add to the literature on how learnable entrepreneurial skills are. Liang

et al. (2018) build a model where individuals learn business skills while working in high paid occupations as workers and show evidence that indirectly confirms the model predictions. Guiso et al. (2021b) show how individuals that grew up in areas with a high density of firms acquire skills useful to run a business. Similarly, we show that individuals who in their last high school year, while being good in math, were exposed to an environment where they could better learn communication skills acquired abilities useful to entrepreneurship.
The paper most close in spirit to ours is Mertz et al. (2023). Mertz et al. (2023) study the effect of early exposure to entrepreneurship on reducing the gap in self-selection into entrepreneurship between men and women. While asking a different research question than the one we address here, their identification strategy and data are similar. Mertz et al. (2023) use within school across cohort variations in the share of female peers' parents that are entrepreneurs or C level managers to study the impact of early exposure to entrepreneurial environments for incentivizing entrepreneurship among women. Our identification strategy also exploits within school across cohort variation, but we use the share of peers' parents graduated in humanities as treatment. Mertz et al. (2023) also have access to danish administrative data, but while they define entrepreneurs as owners of unnicorporated businesses, we define an entrepreneur as the owner of an incorporated business which is active.

4.4 The Data

Our analysis is based on the full population administrative data from Denmark, covering the years from 1996 to 2019. The final dataset is composed of two underlying blocks: the entrepreneurial and the education data set. In the next two sections we describe how we construct these two datasets in detail.

4.4.1 The Entrepreneurial Dataset

The entrepreneurial dataset combines multiple administrative data sources to generate a unique dataset that contains all firm ownership in Denmark between 1996 and 2019. Specifically, by combining individual level characteristics from Statistic Denmark Research Database (DST) with firm level data from the Danish Central Business Register (CVR) and the commercially available KOB database (KOB) from Experian Denmark, we are able to link individual level information to entrepreneurial spells and subsequent business outcomes.

The data contained in Statistics Denmark is provided and updated regularly by relevant Danish authorities, including the Ministry of Taxation, the Ministry of Education and the Ministry of Employment. The database contains general information on individuals such as gender, age, education, wealth and income composition. In addition, detailed employment registers provide all current and previous employment relationships (employer-employee), with corresponding salaries, hours worked, and occupational codes (isco 08) that are used to characterize individual labor market histories, as well as firm-level employment. However, the DST does not contain data on incorporated firms (limited liability companies), but only data on unincorporated firms (sole proprietorship and partnership). As shown by Levine and Rubinstein (2017b), when studying entrepreneurship it is key to separate between owners of sole proprietorships and owners of limited liability companies, as they display very different characteristics. To this end, we add the CVR database to the DST dataset, where the former contains information on all firms registered in Denmark since 1980. The CVR also contains detailed ownership records of sole proprietorships, partnerships and corporations and provides the timing, identity and ownership shares of all direct owners. As ownership records referring to incorporated businesses are limited to the period after 2014, we combine the CVR database with data from the commercially available KOB database, published by Experian Denmark, that contains hand-collected ownership information, which completes missing ownership in the early data years of the CVR database. The KOB database also contains detailed accounting records of corporations. All firms in the resulting dataset are identified by unique CVR-numbers, and all individuals are identified by unique PNR-numbers, which can be matched directly to other data sources.

After combining all these datasets we obtain the entrepreneurial dataset, in which the unit of observation is an individual. For every individual and annually for every year between 1996-2019 the final dataset contains information on individuals' income, net wealth, labor market status, hours worked, occupation, whether an individual owns a sole proprietorship, a partnership or a limited liability company and if so the corresponding business outcomes for each year in which the business exits: revenues, assets, number of employees, turnover, dividends and industry in which the business operates. Concerning years before 1996, the dataset contains, in addition to a variable on accumulated labor market experience, individuals' education level (the highest educational attainment) and her type of education. In addition, for each individual the dataset reports demographic characteristics, the place of birth and the place of living. Finally, we are also able to link individuals

with their parents and their siblings through the PNR number, if alive.

4.4.2 The Education Dataset

The second building block of the final dataset, which we use in our analysis, is the education dataset.Statistics Denmark provides education registries for all student cohorts after 1996. Specifically, for every high school graduate the registries contain information on students' grade point average in the last year of high school. In addition, for all subjects attended by a student, the registries report the grades students have achieved in every examination, as well as in every written and oral assessment and take-home project.

The Danish education system was majorly reformed in 2005. In the interest of working with a homogeneous sample in which grades are comparable, we only keep the cohorts who graduate from high school between 1997 to 2004. Up to 2004, the Danish high school system was characterized by two main tracks students could choose from: a mathematical and a linguistic one. A third track existed, the so called higher preparatory examination (HF), which was designed for young adults who had left the educational system.

In our analysis we focus only on students enrolled in the mathematical high school track. The reason for this is that up to 2004, students in Danish high schools could choose to take subjects at three different levels, corresponding to different difficulties and a different number of hours per subject.[2] Clearly, grades obtained in the same subject but at different levels are not comparable as the difficulty of the classes is very different. While all students in both the mathematical and linguistic track take Danish classes at the highest level (level A), this is not the case for math classes. Only in math oriented high schools students take math classes at the highest level.[3] In order to have skill measures that are as homogeneous as possible, we select only students enrolled in math-track high schools. In this way for every student we observe the grades he obtained both in danish and math classes, taken at the highest, and thus comparable, level.

In the next section we provide an overview of the main descriptive statistics of our sample.

[2]The three different levels were level C (level I), level B (level II) and A (level III). We refer the reader to appendix B for a detailed overview of the Danish high school system.

[3]In Denmark between 1997-2004, 80% of students enrolled in math high schools took math classes at the highest level.

4.5 Descriptive Statistics

We define an entrepreneur as a business owner of a limited liability company, who over the sample period has hired at least one employee and whose business displays positive revenues and assets. Following the work by Levine and Rubinstein (2017b) we define entrepreneurs as owners of incorporated businesses and assure that we do not define as entrepreneur an individual who owns businesses which are empty legal boxes with no economic activity. Henceforth, 'entrepreneurs' will refer to individuals meeting these criteria.

Table 4.1 presents descriptive statistics for firms for whom at least one shareholder is defined as entrepreneur according to our definition. In Part A), we provide statistics for the entire universe of Danish firms between 1997 and 2019, while in Part B), we focus exclusively on firms owned by individuals with mathematical HS degrees between 1997 and 2004. As shown in Table 4.1, both groups of firms are composed of real productive firms. The median firm in the universe of Danish LLCs owned by the defined entrepreneurs has almost 2 employees, is 7 years old, generates more than 500,000 euros in revenue, has over 10,000 euros in net income, possesses almost 350,000 euros in assets, and maintains almost 100,000 euros in equity. Regarding firms owned by individuals with mathematical HS diploma, the values for the discussed variables are higher, although the firms are younger. The presence of 0 employees at the 10th percentile and less than 1 employee at the 25th percentile can be attributed to the observation of firms at their legal registration. It takes some time before a firm begins hiring employees; thus, even if an entrepreneur is required to hire at least one full-time employee to be classified as an entrepreneur, the firm may have 0 employees during its initial years. The same applies to assets and revenue. As for firm net income and dividends (Net Inc. and Div, respectively), they exhibit right-skewed distributions, leading to a median value significantly lower than the mean. The descriptive statistics demonstrate that our criteria for identifying entrepreneurs effectively exclude legal box firms and single-man ventures/freelancers with LLC legal status.

In Table 4.2, we present the distribution of equity ownership shares for our populations of entrepreneurs. In the case of the universe of Danish entrepreneurs (label *All*), the median shareholder is a sole owner of a company, the 25th percentile of ownership is 0.49 and the 10th percentile is 20 percent. As for the population of mathematical high school graduates between 1997-2004, the median shareholder holds 0.50 equity shares of a firm. Consequently, half of the population of share-

holders consists of majority shareholders, by construction. The 1st quartile of the ownership distribution amounts to 0.23, and the 10th percentile is 0.09. Regarding the proper definition of an entrepreneur, we seek individuals who both invest risky capital in a venture and actively participate in firm decision-making. Concerning shareholders who own 50 percent or more of a firm, they are, by construction, majority shareholders and thus actively involved in firm decision making. For those with less than 0.50 ownership, our plan is to identify those who are majority shareholders, and distinguish between minority shareholders actively involved in firm management as employees, and those who are not. We intend to explore this direction further in the future. At present, we are reassured by the fact that both populations of entrepreneurs are predominantly composed of majority shareholders by construction, while the remaining individuals hold sizable and significant amounts of shares to be considered relevant shareholders.
Table 4.3 shows descriptive statistics for the Danish entrepreneurs. In the upper part we report statistics for the general Danish population of men who are older that eighteen in 2019. We see that entrepreneurship is an infrequent career choice. According to our definition of entrepreneurship, slightly less than 4% of the population become entrepreneurs in their life. Those individuals who become entrepreneurs, have on average more years of education (12.8 years against 12.2), and earned higher wages on the labor market, with a difference of around 19 percentage points.

Selection into entrepreneurship does not happen with the same frequency at different stages of the life-cycle. In Figure 4.1 we plot the probability of becoming an entrepreneur by age for the cohorts born between 1961 to 1975. We see that the probability of becoming an entrepreneur is hump-shaped in age, with a peak around age 36. These life-cycle patterns reveal that we ought to be able to follow individuals for long time periods over their life to study the drivers of selection into entrepreneurship.

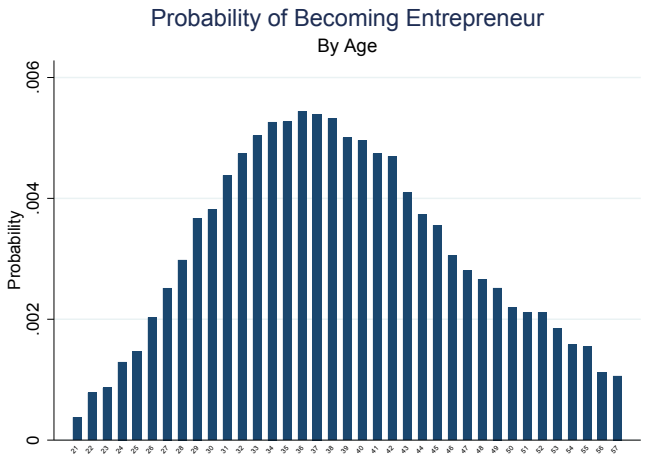
The second block of Table 4.3, with the heading year of birth 1979-1986, shows descriptive statistics for the cohorts for which we observe the high school grades, as described in the previous section. In fact, individuals that graduated from high school between 1997 and 2004 belong to the 1979-1986 cohorts. For this subsample, the share of individuals that become entrepreneurs is the same as for the general population, around 4%. However, we do not find the positive selection in

Table 4.1: Firms Descriptive Statistics

A)	Population of Danish llc: year 1996-2019								
	Emp.	Age	Rev.	Net Inc.	Asset	Equity	Ebit	Div.	V.Add.
10 p.	0	1	959	-23,638	26,260	-34	-17,894	0	7
25 p.	0.24	3	170,561	0	115,209	20,399	0	0	50,963
50 p.	1.85	7	523,124	13,808	346,716	95,531	24,446	6,552	180,960
75 p.	5.54	12.47	1,419,970	61,195	985,459	319,079	92,154	32,500	481,310
90 p.	13.2	18.47	3,714,969	188,745	2,823,350	953,557	266,326	120,871	1,145,667
Mean	7.13	8.34	2,536,569	139,896	3,457,400	1,247,211	200,729	124,730	680,157
B)	Population of Danish LLC Owned by a Math HS Grad.: year 1996-2019								
	Emp.	Age	Rev.	Net Inc.	Asset	Equity	Ebit	Div.	V.Add.
10 p.	0	0	54,018	-46,262	38,361	-5,237	-41,412	0	7,102
25 p.	.65	1.48	229,198	0	130,531	20,425	209	0	77,106
50 p.	2.79	4	683,672	23,836	427,789	108,187	35,050	12,165	258,294
75 p.	7.99	7	2,070,924	112,964	1,401,076	431,254	154,789	70,196	749,924
90 p.	20.81	10.46	5,725,121	407,232	5,197,500	1,656,807	523,599	263,180	1,981,345
Mean	11.5	4.76	3,863,476	264,987	5,187,148	1,807,885	343,241	207,360	1,133,255

Notes — The table reports descriptive statistics for firm outcomes and characteristics owned by entrepreneurs who meet our criteria for being defined as entrepreneurs. A) Danish llc population of firms between 1997 to 2019 B) Danish llc population of firms between 1997 to 2019, owned by math HS graduates between 1997 to 2004. Values are in real euro 2015. "Emp" stands for number of firm full time employees, and "Age" for firm age. Percentile are computed as average around percentile to comply with data regulation.

Figure 4.1: Probability of becoming entrepreneur by age



Notes: Share of individuals that become entrepreneurs at a given age in the Danish population of men for the cohorts 1961-1975. The data starts in 1996 and ends in 2019.

terms of education and wages, which we observed in the general population. Part of the explanation is that entrepreneurship is a career choice undertaken late in life, as shown in Figure 4.1. This implies that individuals of cohorts 1979-1986 who are pursuing university degrees, have not yet had the time to actually become en-

Table 4.2: Equity Shares of Entrepreneurs

	10 p.	25 p.	Median	75 p.	90 p.	Mean
All	.20	.49	1	1	1	.69
Math HS 1997-2004	.09	.23	.50	.99	1	.49

Notes — The table reports descriptive statistics of equity share firm ownership among entrepreneurs. 'All' refers to the universe of Danish entrepreneurs, and 'Math HS 1997-2004' to the universe of Danish entrepreneurs who graduated from a mathematical high school between 1997 and 2004. Firm ownership period of observation: 1997-2019. Percentile are computed as average around percentile to comply with data regulation.

trepreneurs. This explains why in the subsample 1979-1986, future entrepreneurs are slightly less educated. Similarly, the negative selection of future entrepreneurs in terms of wages likely stems from the fact that we are comparing wages of future entrepreneurs when they are relatively young, with wages of individuals who always remain workers and thus are older on average[4]. Given that wages increase on over the life-cycle, this explains why on average future entrepreneurs display lower earnings than always workers.

Of the 1979-1986 cohorts, 37% have a high school degree (gymnasiale uddanelser), while the rest do not. Individuals who have a high school diploma are likelier to become entrepreneurs (4.15% against 3.55%) and unsurprisingly have more years of education and earn more. This can be seen in Table 4.4.

We further provide statistics for the subsample of individuals who completed their high school degree (gymnasiale uddanelser) between the years 1997 and 2004 in Table 4.5. Among the 61 036 men completing high school during these years, 64% of them completed a mathematical high-school program (39 270 individuals), 21% of them a linguistic one (12 827 individuals) and15% of them the higher preparatory examination high school (8 939 individuals). The absolute number of individuals enrolled in the different high school tracks, with respect to the overall number of high school graduates, is the result of different gender ratios in the three tracks. The mathematical high school program is gender balanced with a 50% of men and women, the linguistic program is female dominated with only 22% of men, and similarly for the higher preparatory examination school with a ratio of men to women of 0.3. Mathematical high school students dominate the other students in terms of years of education, gpa and wages when working. A clear hierarchy emerges where mathematical high school students are more skilled than

[4]By definition wages are only observed for paid employed workers.

Table 4.3: Summary Statistics A

	Danish entrepreneurs	Rest of the population
Men older than 18		
Absolute number	110,356	2,745,731
Share	3.86%	96.14%
Average years of Education	12.78	12.15
	(.007)	(.001)
Log wage	5.50	5.31
	(.0006)	(.0000)
Year of birth 1979-1986		
Absolute number	11,162	285,267
Share	3.77 %	96.23%
Average years of Education	12.86	13.20
	(.02)	(.005)
Average log wage	5.08	5.10
	(.001)	(.0002)
Mathematical HS Students:Year of Graduation 1997-2004		
Absolute number	1,885	37,385
Share	4.8%	95.2%
Average Years of Education	15.21	15.71
	(.05)	(.01)
Average log wages	5.24	5.23
	(.004)	(.000)
Average log GPA	4.18	4.20
	(.009)	(.002)
Average grade in Danish	8.52	8.60
	(.05)	(.01)
Average grade in Written Danish	8.31	8.41
	(.05)	(.01)
Average Grade in Oral Danish	8.77	8.81
	(.06)	(.01)
Average Grade in Math	7.82	8.12
	(.08)	(.019)

Notes — This table reports summary statistics of the data used in the analysis.

linguistic high school ones, which in turn are more skilled than higher preparatory examination graduates.

In Appendix A, we display the kernel distributions of GPA (Figure C.1), grade in math (Figure C.2), and Danish (Figure C.3) during the final year of high school for men graduating from *Gymnasiale uddannelser* (upper secondary education programs) between the years 1997 and 2004.

The final block of Table 4.3 displays descriptive statistics for the sample we work with, namely the universe of danish male individuals who graduated from a mathematical high school track between 1997 and 2004. Their average age in 2019 -

Table 4.4: Summary Statistics B

	Gymnasiale Graduates	Non Gymnasiale Graduates
Year of birth 1979-1986		
Number	106,960	189,469
Share	36.08%	63.92%
Share of Entrepreneurs	4.15%	3.55%
Average Years of Education	14.7	12.3
	(.007)	(.006)
Average log wage	5.13	5.09
	(.000)	(.000)

Notes — This table reports summary statistics of the data used in the analysis.

Table 4.5: Summary Statistics C

	Mathematical HS	Linguistic HS	HF
Entire population			
Absolute number	39,270	12,827	8,939
Shares	64%	21%	15%
Gender ratio(male)	0.50	0.22	0.30
Average Years of Education	15.69	15.04	14.08
	(.01)	(.02)	(.02)
Average log wage	5.23	5.15	5.14
	(0.00)	(.001)	(.001)
Average log GPA	4.20	4.14	3.99
	(.002)	(.003)	(.005)
Share Level A Danish	100%	100%	100%
Share Level A Math	80%	0.0%	7%

Notes — This table reports summary statistics of the data used in the analysis.

the last year of the sample - is 38.3 and the median individual is born in 1981. Approximately 4.8% of individuals become entrepreneurs, 1885 out of the 39 270 individuals in total. In this subsample entrepreneurs are slightly less educated in terms of years of education and earned on average 1 percentage point higher wages compared to individuals who never start a business. When compared at the same age, future entrepreneurs earned 10 percentage points higher wages compared to always workers. Regarding educational outcomes we find that entrepreneurs have on average slightly lower gpa (2 p.p less) and also display slightly lower grades in all the different types of danish examinations. In math, future entrepreneurs report an average grade which is about 5% less than the average math grade of individuals who never start a business.

In the next section, we examine the relationship between schooling, labor market

outcomes, and selection into entrepreneurship.

4.6 Balanced skills and labor market outcomes

Motivated by the theory of Lazear (2004b), in this section we study how different compositions of skill sets translate into labor market outcomes for paid employed workers and how they relate to selection into entrepreneurship. We use high school grades individuals obtained in danish to capture communication skills, grades obtained in math classes to measure analytical skills and their interaction to capture skill multidimensionality. In the first subsection we provide simple correlational evidence on the association between high school performance and labor market outcomes. In the second subsection we move to the relationship between schooling performance and entrepreneurship, and in the third subsection we deepen our understanding of the complementarity between analytical and communication skills for self-selection into entrepreneurship.

4.6.1 Returns to schooling on the labor market

Table 4.6: High school grades and labor market outcomes

	(2) Log-wage	(3) Log-wage	(4) Log-wage	(5) Log-wage
log_gpa	0.120***			
	(0.00199)			
bStdX	0.044			
Danish		0.00354***		
		(0.000369)		
bStdX		0.008		
Math		0.0113***	0.0117***	0.0131***
		(0.000223)	(0.000216)	(0.000639)
bStdX		0.041	0.042	0.048
Danish Oral			0.00206***	0.00368***
			(0.000287)	(0.000716)
bStdX			0.006	0.010
Danish_Oral#Math				−0.000169*
				(0.0000681)
				-0.008
Constant	4.731***	5.114***	5.123***	5.109***
	(0.00845)	(0.00298)	(0.00261)	(0.00612)
N	561674	563904	563904	563904

Notes: * (p¡0.1), ** (p¡0.05), ***(p¡0.01). Sample of Mathematical high school students (HS) that graduate from HS between 1997 to 2004 and attended level A math courses. Log-wage is the logarithm of real-wage. bStDX stands for the variation in the expected value of the outcome variable for a standard deviation variation in the explanatory variable.

In Table 4.6 we report the returns on the labor market to the final H.S. year GPA, the average grade in math and danish for the sample of mathematical high school students graduated between 1997 and 2004. As we can see, a one standard deviation increase in log gpa increases expected real wages on the labor market by 4.4 percentage points, or, otherwise said, a 1% higher gpa predicts a 0.12 % higher real wage. In Appendix C, we display the difference in log-wages for each decile of the GPA distribution, showing an increasing linear pattern. When looking at the grades in math and danish we find that mathematical skills are more rewarded. A one standard deviation higher average grade in math predicts an expected real wage of 4.1 percentage points higher, while the same increases in the average grade in danish is only associated with 0.8 percentage points higher wages. The same patterns holds true if we look at grades in only the oral examinations of danish.
In column 5 of Table 4.6, we interact math and danish grades to explore whether individuals with more balanced skill sets receive a premium on the labor market as paid employed workers. We find that the coefficient for the interaction term is negative, suggesting that the labor market seems to favor individuals with specialized, rather than balanced skills. This evidence seems to suggest that the labor market rewards skill specialization.

4.6.2 High school grades and selection into entrepreneurship

So far we have established a positive association between average school performance, as measured by log gpa, and future wages as paid employed workers. Moreover, we have established that the labor market rewards specialized workers more than multidimensional ones.
In this section we ask how school outcomes and the complementarity between different skills relate to the probability of selection into entrepreneurship. We start by asking how average schooling ability, as measure by log gpa, is associated with the decision to start a business. We find that log gpa is negatively correlated with the probability of selecting into entrepreneurship. In the first column of Table 4.7 we see that a one standard deviation higher log gpa decreases the probability of becoming entrepreneur by 0.4 percentage points, which is 10% of the average share of entrepreneurs in the sample. In Figure 2, we illustrate the difference in the average share of entrepreneurs for each decile relative to the first decile of GPA. We observe that the negative relationship between GPA and the probability to become an entrepreneur is a result of the fact that individuals in the first decile of GPA

exhibit the highest share of entrepreneurs, while the pattern for deciles higher than the first is flat. Next, we study how communication and analytical skills and their complementarity are associated with the probability of becoming an entrepreneur. As before, we use grades obtained in danish and math to measure the two different skills. In Table 4.7 we report the results for the following OLS regression:

$$Ent_i = \alpha + \beta_1 \times Math_i + \beta_2 \times Danish_i + \beta_3 \times (Math_i \times Danish_i) + \sum_{y=1997}^{2004} \gamma_y * Year + \varepsilon_i \tag{4.1}$$

where *Ent* is 1 if an individual ever becomes an entrepreneur over the sample period and 0 otherwise. *Math* and *Danish* are the average grades of all the math and danish grades received by the individual during the last year of high school and *Year* are year fixed effects for the final H.S. year. This specification describes the association between the grades in danish and math received in high school by the student, and their interaction, given the year in which students graduate. The year fixed effects control for the fact that later cohorts have less time to become entrepreneurs in their life as the sample period ends in 2019. They also control for possible "grade inflation" that can affect the grading system across cohorts. We cluster standard errors at the year-school level.

The table shows three different specifications. In the second column of Table 4.7 *Danish* is the average of every grade obtained across all types of examinations in danish. In column (3) we only use grades obtained in oral danish exams, while in column (4) we only use grades obtained in written evaluations of danish. The grade in *Math*, instead, is always the mean across all types of examination in math, both written and oral. The second specification shows that: i) there is a negative, but not significant, association between the probability of ever becoming an entrepreneur and the average grade in danish ii) a negative and significant association between ever becoming an entrepreneur and the average grade in math iii) a positive, but slightly insignificant association between the interaction term and the probability of becoming an entrepreneur. Specifications (3) and (4) help us understand whether oral or written danish skills drive the above relationship. We see that further splitting performances in danish into oral and written skills changes our results. Specifically, once we consider only oral danish evaluations the association between the probability of selecting into entrepreneurship and the interaction term becomes more significant. When, instead, we only consider written evaluations in danish, as in specification (4), the interaction term goes down to zero and becomes

insignificant. This suggests that oral danish skills, interacted with math skills, are the ones predicting selection into entrepreneurship. The magnitudes of these associations are sizable. From specification (3) we see that a one standard deviation increase in math is associated with a decrease in the probability of becoming an entrepreneur of 1.3 percentage points, and a one standard deviation increase in the interaction term between oral and math skills increases the probability of becoming an entrepreneur by 1.3 percentage points. These numbers are economically relevant if we consider that the overall share of entrepreneurs in the sample is 4.9%.

To further gain understanding of the magnitudes of the coefficients let us consider two different individuals who have very different skill set compositions. Let us consider a first individual with a perfectly balanced skill set, who has grade 8 both in danish and math and a second individual who has an extremely specialized set of skills with grade 15 in math and 1 in danish.[5] Using the estimated coefficients from specification (3) this would imply that the individual with a specialized skill set- grade 15 in math and 1 in danish- has a probability of selecting into entrepreneurship of 3.7%, while the individual with a balanced skill set- grade 8 both in math and danish- has almost a double probability of 6.7%.

In Figure 4.3, we display the standardized coefficient β_3 from equation 4.1, interacted with the age of the individual. The coefficients shown in Figure 4.3 for each age represent the variation in the shares of entrepreneurs at a given age for a one-standard-deviation change in the product of the math and oral grades. As can be seen in the figure, the relationship increases with age, consistent with the fact that entrepreneurship is a decision undertaken later in life. Particularly, the association is null at a very young age (up to 28 years old) when individuals do not open firms, and starts to increase significantly in the late 30s when the probability of becoming an entrepreneur is at its maximum (Figure 4.1).

Together, this provides observational evidence that multidimensional skill sets predict selection into entrepreneurship and that different compositions of skills have meaningful economic significance to understand the decision to start a business. Moreover, our evidence shows that the oral examinations in danish seem to best capture communication skills.

[5]The Danish grading system goes from -3 to 12. We converted it to a 1-16 scale to facilitate the interpretation. A grade of 8 is considered a fair grade, 15 is considered an excellent grade while 1 is unacceptable. The appendix contains detailed information on the danish high school grading system.

Table 4.7: High school grades and selection into entrepreneurship

	(-)	(-)	Danish Oral	Danish Written
	(1)	(2)	(3)	(4)
	Entrepreneur	Entrepreneur	Entrepreneur	Entrepreneur
log_gpa	−.0099***			
	(.003)			
bStdX	-0.004			
Danish		-0.00224	-0.00207	-0.000903
		(0.00166)	(0.00138)	(0.00149)
bStdX		-0.005	-0.006	-0.002
Math		−0.00337**	−0.00370***	-0.00202
		(0.00135)	(0.00116)	(0.00123)
bStdX		-0.012	-0.013	-0.007
Danish×Math		0.000237	0.000257**	0.0000863
		(0.000150)	(0.000126)	(0.000138)
bStdX		0.011	0.013	0.004
Graduation year f.e.		✓	✓	✓
N		31293	31293	31292

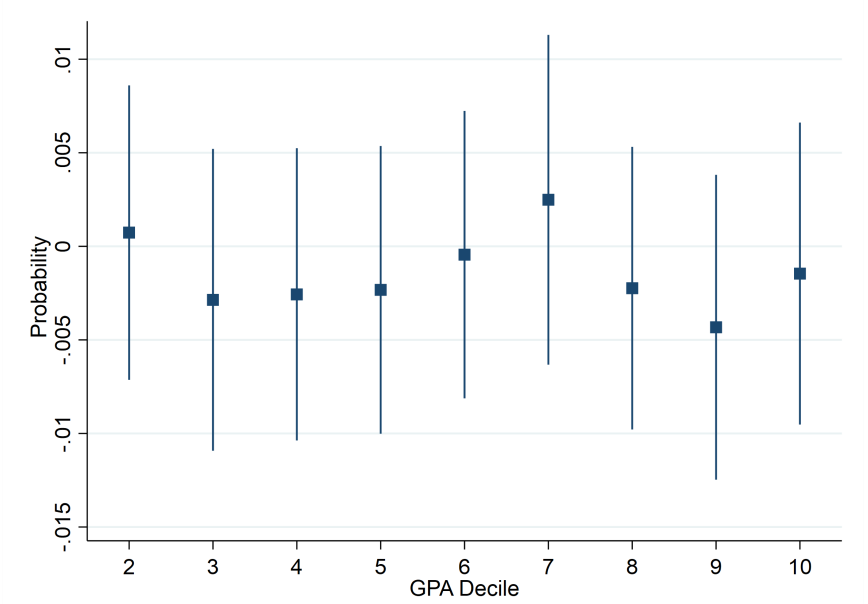
Notes: * (p¡0.1), ** (p¡0.05), *** (p¡0.01). The table reports OLS coefficients of the regression of the probability of being an entrepreneur on log-gpa and HS school grades in math, danish, and their interaction. The sample is the universe of male Danish Mathematical high school students in their last HS school year that attended were enrolled in the last high school year between 1997 to 2004. The outcome variable *Entrepreneur* is 1 if an individual has ever been an entrepreneur in his life. The explanatory variable *Danish* is in specification (2) the average of the grades received in Danish, in (3) the average grades received in solely the oral evaluation of Danish, while in (4) the average grades received solely in the written evaluation of Danish. The explanatory variable *Math* is the average of the grades received in math courses. Grades in Danish and Math are all for level A course. The regression contains additional controls for graduation year f.e.. Standard errors are clustered at school-year-programme in parentheses: # 1113calendar year.

4.6.3 Teach the nerds to give a pitch: Oral skills and Very Good Math students

In this section we further deepen our understanding of the relationship between skill complementarity and selection into entrepreneurship. We do this by grouping students into three categories based on their performance in danish and math. Specifically, using official definition of grades from the Danish Ministry of Higher Education and Science we define three grade categories and assign individuals based on their mean grade in each subject, computed over all the grades in a given subject obtained in the last year of high school. [6] The categories are the same for danish and math and are: *Bad*, *Average* and *Very Good*. Category *Bad* contains all students with a mean grade that is less than 4 (8 according to our scale). These

[6]We refer the reader to the appendix for an extended definition of grades in Denmark. The source for the Danish Grading system can be found at: https://ufm.dk/en/education/the-danish-education-system/grading-system

Figure 4.2: GPA Deciles and Probability of Becoming Entrepreneur



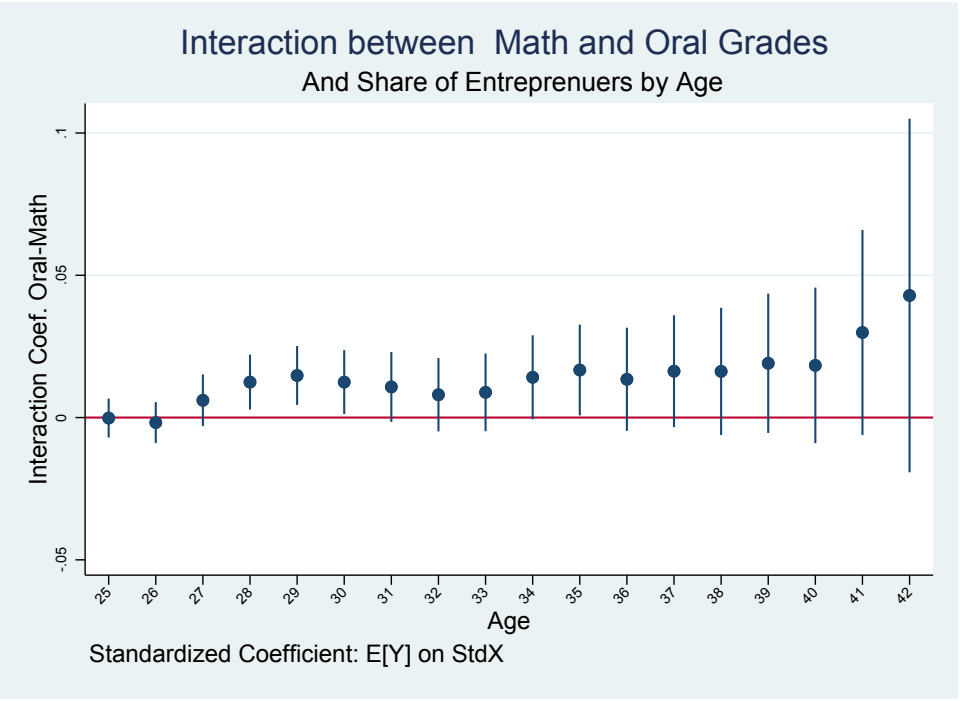
Notes— The figure displays the coefficients for GPA decile dummies from a regression of the probability of becoming an entrepreneur, during the observation period, on GPA decile and year of bith f.e. . The coefficients show the difference in average share of entrepreneurs between a decile and the first decile. Population: H.S. graduates between 1997 and 2004. Period of observation: 1997 to 2019.

are students whose performance is either unacceptable, does not meet the minimum requirements, or only meets the minimum requirement and is below a fair evaluation. The category *Average* contains all individuals with an average grade greater or equal than 4 (8 according to our scale) and less or equal than 7 (11 according to our scale). These are students whose performance is fair, between fair and good, and good. The final category *Very Good* is composed of all students with an average grade higher than 7 (11 according to our scale), who are students with a performance that is more than good, very good or excellent.
Dividing students into these categories helps us gauge more insights on the role of skill complementarity. We do this by running the following regression:

$$Ent_i = \alpha + \beta_1 \times Average_i + \beta_2 \times VeryGood_i + \sum_{y=1997}^{2004} \gamma_y * Year + \varepsilon_i \qquad (4.2)$$

Where *Average* and *Very Good* are dummies taking value 1 when an individual has a mean grade falling into one of these categories. The variable *Year* are year fixed

Figure 4.3: Interaction coefficient between oral and math grade by age: standardized coefficients



Notes— The figure displays the coefficient β_3 from regression 4.1 in a specification where $Math_i \times Danish_i$ is interacted with the age of the individual. Coefficients are standardized and report the variation in the probability of being an entrepreneur at every age for a one standard deviation variation in the product of the oral and math grades. Population: H.S. math graduates between 1997 and 2004. Period of observation: 1997 to 2019.

effects for the year in which the individual was graded. We cluster standard errors at the year-school level. With this specification *Average* and *Very Good* read as the difference in the share of entrepreneurs in those categories with respect to the base category *Bad*.

In table 4.8 we report the regression coefficients of equation 4.2. In column (1) of Table 4.8 we subset for the individuals that are *Bad* in math and check how the probability of becoming an entrepreneur changes as we move along the categories of the danish grades, from *Bad* to *Very Good*. We see that there are no significant differences in the share of entrepreneurs as we move from the group of *Bad* (the baseline) students to the group of *Very Good* in the oral danish exams. We observe the same qualitative pattern in column (2) where we subset for the students that belong to the *Average* category in math.

On the contrary, when we look at the relationship between danish oral skills and

Table 4.8: Probability of becoming entrepreneur by skills

	Average Grade in Math		
	Bad Entrepreneur	Average Entrepreneur	Very Good Entrepreneur
Oral Danish Average	-0.00540	-0.00330	0.0182**
	(0.00475)	(0.00618)	(0.00821)
Oral Danish Very Good	-0.00307	0.000737	0.0259***
	(0.00956)	(0.00733)	(0.00882)
Constant	0.0679***	0.0593***	0.0447***
	(0.00752)	(0.00741)	(0.00972)
N	11790	12345	7158

Notes: * (p¡0.1), ** (p¡0.05), *** (p¡0.01). This table reports OLS coefficients of the regression of the probability of being an entrepreneur on HS school grades in Danish Oral examination for different levels of the grades in math. The sample is the universe of male Danish Mathematical high school students in their last HS school year attended between 1997 and 2004 . Specification (1) displays the difference in the share of entrepreneurs by average grade in oral Danish for students that have a bad average grade in Math. Specification (2) displays the difference in the share of entrepreneurs by average grade in oral Danish for students that have an average average in Math. Specification (3) displays the difference in the share of entrepreneurs by average grade in oral Danish for students that have a very good average grade in Math. The baseline is the average share of entrepreneurs for students that have a Bad average grade in Danish Oral. Grades in Danish and Math are all for level A course. The regression contains additional controls for graduation year f.e. Standard errors are clustered at school-year-program in parentheses: (1) # 1104, (2) #1090, (3) 1070

the probability of selection into entrepreneurship for students that are very good in math, we observe an increasing pattern as we move along the different categories of danish oral exams. Very good math students who score *Average* in danish oral have a 1.8 percentage points higher probability of becoming an entrepreneur compared to the baseline of *Bad* students. In turn, very good math students who belong to the *Very Good* group in oral danish exams have 2.6 percentage points higher probability of ever becoming entrepreneurs compared to the baseline category. Considering that for the baseline group of *Bad* students the probability of becoming entrepreneurs is 4.5%, it means that high ability math students scoring very well in danish oral exams have a 60% higher probability of transitioning into entrepreneurship compared to talented math students that score poorly on danish.
In Table 4.9 below we report the shares of entrepreneurs for the different combinations of groups (9 combinations in total), without controlling for the high school year of graduation (standard error in parenthesis). From the table we observe another couple of facts. As already found with regression 4.2, for students who are very good in math the share of individuals that become entrepreneurs is strongly increasing in the oral score in danish. This pattern, however, is not present for stu-

dents who are bad or average in math. Second, the highest share of entrepreneurs is among the students that are bad both in oral danish exams and math.
This two findings align with evidence from the previous sections. In particular, the last finding is in accordance with log gpa being negatively associated with selection into entrepreneurship..
The most interesting correlational evidence is that the probability of selecting into entrepreneurship for high skilled math students is strongly increasing in oral danish abilities. This finding can be rationalized through models of entrepreneurship that build on the intuition initially proposed by Lazear (2004b), for which individuals that start businesses need to be able to perform a variety of different tasks and must thus be multidimensional in their skill set. More recent research by Choi et al. (2019), who analyze the human capital composition of founding teams in start-ups, seem to give similar importance to the multidimensionality of human capital for the understanding of business outcomes.

Table 4.9: Share of entrepreneurs by group of grades

	Math grade		
	Bad	Average	Very good
Danish grade			
Bad	0.057	0.052	0.023
	(0.004)	(0.006)	(0.008)
Average	0.052	0.048	0.039
	(0.002)	(0.002)	(0.003)
Very good	0.054	0.051	0.046
	(0.009)	(0.005)	(0.004)

Notes: The table reports the share of entrepreneurs by group of grades without any additional controls.

4.7 Entrepreneurial outcomes of math skilled students

In this section we study the performance of businesses owned by individuals who attended a mathematical high school and took math classes at the highest level. We are motivated by our previous findings that for the group of talented math students the share of entrepreneurs increases as their communication skills improve. We now want to know whether firms owned by individuals with high mathematical skills on average also generate more revenues, more employment and are more profitable. If this is the case, then asking how we can incentivize entrepreneurship among students with high analytical skills becomes a policy relevant question to

investigate.

To simply the analysis of this section, whenever in our data we have individuals who are owners of multiple businesses we only keep the outcomes of the firm in which the individual holds the highest share. This helps us create a more direct link between individuals and firm performance. We start by running a simple regression in which we study the association between attending a mathematical high school - and taking level A math classes- and firm performance. We run the following regression:

$$Out_{it} = \alpha + \beta_1 * MathHS_i + \sum_{y=1997}^{2004} \gamma_y * Yearbirth + \sum_g \theta_g * year + \varepsilon_{it} \qquad (4.3)$$

where Out_{it} is firm outcome in year t, for the firm owned by individual i, $MathHS_i$ is a dummy taking value one if the entrepreneur attended a mathematical high school and *Yearbirth* and *year* are respectively the year of birth of the entrepreneur and year fixed effects. The coefficient *MathHS* displays the average difference in firms outcome with respect to the general population of firms owned by individuals born between 1979 and 1985, for firms that are owned by an individual who attended a mathematical high school and took math courses at the highest level (level A). [7] We control for the year of birth of the owner to control for the longer period older individuals have to open and manage a firm, and we control for year fixed effects to account for aggregate economic conditions. The unit of observation is the firm-year outcome.

Table 4.10 reports the coefficients of regression 4.3. Starting from the upper left column we see that firms owned by individuals who attended mathematical high schools have: i) a higher number of employees- 1 more employee with respect to the average of the general population which is approximately 8, that is 12,5% higher number of employees compared to the baseline; ii) almost 7 percentage points higher revenue; iii) 17 percentage points higher value added; iv) 26 percentage points higher assets; v) 32 percentage points higher earnings before interests and taxes (Ebit); vi) 35 percentage points higher net income; viii) 34 percentage points higher value of equity. To sum up, these means that an entrepreneur who

[7] The set of observations is composed of all firm outcomes of businesses owned by individuals born between 1979 and 1985 in order to be comparable to firm outcomes of firms owned by individuals who attended a mathematical high school, as the education dataset is available for cohorts graduated between1997 and 2004

attended a mathematical high school on average holds bigger firms (in terms of employment, revenues and assets) and more profitable firms (ebit and net income), compared to entrepreneurs who did not attend a mathematical high school.

Table 4.10: Firm outcomes of mathematical high school students

	Employees	Log revenues	Log value added	Log assets	Log Ebit	Log net income	Log equity
Math HS	1.092***	0.0692***	0.175***	0.269***	0.337***	0.361***	0.351***
	(0.124)	(0.0160)	(0.0154)	(0.0179)	(0.0241)	(0.0254)	(0.0223)
Constant	7.477**	8.335***	7.827***	8.469***	6.286***	5.338***	7.397***
	(2.883)	(0.369)	(0.330)	(0.390)	(0.561)	(0.538)	(0.451)
N	45741	45274	44752	45104	36623	35004	39768

Notes: * $(p < 0.1)$, ** $(p < 0.05)$, *** $(p < 0.01)$. This table reports OLS coefficient of the regression of firm outcomes on having been in Mathematical HS. The coefficient Math HS measures the average difference in firms outcomes owned by students who attended a mathematical high school, compared to outcomes of firms owned by students who attended other high schools. An individual is considered attending a Mathematical HS if she attended Mathematical HS and took level A math classes (80% of the sample). The sample is the universe of all limited liability companies having at least one employee owned by an individual born between 1979 and 1986. The regression contains additional controls for year of birth of the owner and year fixed effect. All outcome variables are defined as in the Danish Authority accounting standards, apart from employment they all log, and they are all trimmed at the 1st and 99th percentile. In terms of entrepreneurs the sample is composed by 11162 entrepreneurs, of which 1392 are mathematical HS.

In Table 4.11 we report the coefficients of running the same regression as before, but in which we split individuals according to our three categories of *Bad*, *Average* and *Very Good* math skills.[8] Thus the coefficients *Bad Math*, *Average Math*, *Very Good Math* show the difference in average firm outcomes for firm owned by individuals that attended a math high school with an average math grade being either *Bad*, *Average* or *Very Good*, with respect to the general population of entrepreneurs born between 1979 and 1985. We find that the better an individual was in math at high school, the bigger and the more profitable his firm is. In terms of profitability, an entrepreneur who belonged to the *Very Good* math group in high school owns firms that have an ebit of 54 percentage points higher than the general population, while entrepreneurs who belong to the *Bad* math category have businesses with an ebit which is only 22 percentage points higher than the baseline. Similarly for net income, where *Very Good* math business owners have firms which display net income that is 55 percentage points higher than the general population, while for the *Bad* group it is only 24 percentage points higher. Also in terms of revenues and employment we see that i) individuals who belong to the *Very Good* category in math skills own firms that have 2 employees more than the

[8]The three categories are defined as in the previous section.

rest the population (25% more), while individuals who belong to the *Bad* group own firms that have the same amount of employees as the general population; ii) *Very Good* math entrepreneurs also have firms that on average have 16 percentage points higher revenue, while *Bad* math entrepreneurs only 5 percentage points higher. These findings hold true also for the other firm performance measures as can be seen in Table 4.11.

Table 4.11: High school grades and firm outcomes

	Employees	Log revenues	Log value added	Log assets	Log Ebit	Log net income	Log equity
Bad_Math	0.311	0.0525*	0.0508*	0.180***	0.242***	0.252***	0.215***
	(0.190)	(0.0243)	(0.0233)	(0.0273)	(0.0369)	(0.0390)	(0.0341)
Average_Math	1.442***	0.0428	0.218***	0.255***	0.334***	0.378***	0.337***
	(0.188)	(0.0244)	(0.0234)	(0.0273)	(0.0363)	(0.0382)	(0.0338)
Very_Good_Math	1.957***	0.159***	0.348***	0.482***	0.540***	0.549***	0.649***
	(0.267)	(0.0348)	(0.0334)	(0.0386)	(0.0524)	(0.0550)	(0.0474)
_cons	7.531**	8.341***	7.836***	8.481***	6.288***	5.352***	7.415***
	(2.883)	(0.369)	(0.329)	(0.390)	(0.560)	(0.538)	(0.451)
N	45741	45274	44752	45104	36623	35004	39768

Notes: * (p¡0.1), ** (p¡0.05), *** (p¡0.01). This table reports OLS coefficient of the regression of firm outcomes on having been in a mathematical HS and having a Bad, Average, very Good grade average in Math. An individual is considered attending Mathematical HS if she attended Mathematical HS and took level A math (80% of the sample).The sample is the universe of all limited liability companies having at least one employee owned by an individual born between 1979 and 1986. The regression contains additional controls for year of birth of the owner and year fixed effect. All outcome variables are defined as in the Danish Authority accounting standards, apart from employment they all log, and they are all trimmed at the 1st and 99th percentile. In terms of entrepreneurs the sample we have 11162 entrepreneurs, of which 1392 are mathematical HS, 471 are Bad_Math, 446 are Average_Math, and 210 Very_Good_Math.

To sum up, this evidence shows that individuals who attended a mathematical high school program own firms that are bigger and more profitable compared to the rest of the population and that both firm size and profitability is increasing with analytical skills, as measured by math grades in high school. While suggestive, these findings point towards the idea that quantifying the role of complementarity between communication and analytical skills in driving self-selection into entrepreneurship for the group of high skilled math students is a policy relevant question to explore in light of the above average quality of firms that math skilled entrepreneurs run. This leads us to the next section in which we propose an identification strategy to casually estimate and quantify the effect of increasing communication skills for mathematical high school students on the probability of starting a business.

4.8 Identification Strategy

So far we have provided compelling observational evidence on the relationship between individuals' skill set composition and selection into entrepreneurship. On a correlational level we observe that i) individuals who were good at math in high school run more successful and profitable businesses ii) the share of good math high school students who start a business is increasing in their communication skills. In this section we want to do an additional step and casually estimate the effect of improving communication skills of high ability math students on the probability that they will become entrepreneurs. Quantifying the effect of this treatment for the population of high skilled math students is crucial if policymakers want to design training programs or other policy interventions to incentivize the creation of high performing firms.

To causally estimate the effect of communication skills on selection into entrepreneurship for the population of high skilled mathematical high school students, we draw from the literature on peer effects. The general idea of this literature is that individuals learn from each others and that skills get transferred across students in the same school. We use information on parental education and exploit within school, across cohort variation in the share of parents peer students' with an academic background in humanities (Mertz et al. (2023)). Motivated by the fact that human capital and skills get transmitted across generations, the basic idea is to instrument individual communication skills with the human capital composition of parents peers' classmates. In particular, for every individual who attended a math high school and took level A math classes we compute the share of his parents schoolmates' who have a university degree in the field of humanities[9]. We then regress a dummy that takes value one if an individual ever becomes an entrepreneur in our sample on the share of fathers peer students' graduated in humanities, controlling for school fixed effects and school fixed effects interacted with a variables that measures the share of the parents of the peer students across the different possible education levels[10]. In other words, for the seven cohorts (1997-2004) of students graduated in mathematical high schools we compare individuals who in their last year of high school were in the same school, in two different cohorts -

[9] We consider university degrees that take 5 years to be completed (so BSc + Msc).

[10] This means that the variable measures the share of parents peers' that have 9,12,15,17 or 20 years of education, corresponding to an education level of less than high school, high school, BSc, MSc and PhD.

but exposed to the same share of fathers peers' at every education level- where for one student the fathers of his peers are more frequently graduated in humanities than other fields. We additionally control for the education level and the area of education of the father of every individual as we adopt a leave one out strategy. [11]. The specification is:

$$Ent_i = \alpha + \beta_3 * Hum_{J-i,i} + \sum_{sc=n,edccs=k}^{sc=N,edcs=K} \Gamma_{n,k} * Sc_n * edcs_k + \theta * C_i + \varepsilon_i \qquad (4.4)$$

Where Ent_i is a dummy taking value one if the individual ever becomes an entrepreneur in our sample, $Hum_{J-i,i}$ is the share of fathers peers' graduated in humanities during the last high school year of the individual,$Sc_n * edcs_k$ is the cross product between a school fixed effect and the share of maximal education level of fathers' peers, C_i contains the individual's father education level and field of graduation as well as municipality fixed effects. So for two individuals in the final year of high school, who attended the same mathematical high school in different years, whose parents peers' however have the same educational structure (same share of fathers peers' with 9,12,14,15,17,20 years of education), we estimate the difference in the share of entrepreneurs for the individual that was exposed to a higher share of peer students with parents graduated in humanities. This identification strategy has two key assumptions. First, we assume that individuals did not strategically self-select into school programs with schoolmates that were more frequently sons of fathers with university degrees in humanities. The second identifying assumption is that sons of fathers graduated in humanities are on average better in communication skills and that these skills transfer to peer students. This last assumption is testable. We discuss this assumption in section 7.2.

4.8.1 Variability in treatment

In the literature on peer effects, a common concern with our type of identification strategy is whether there is enough variation in peer characteristics. In our setting, this means asking if across cohorts of students that graduated from the same mathematical high school during the years 1997-2004 and who have the same number of fathers with a university degree, we observe enough variation in the share of fathers peers' with a humanistic degree.
We address this concern in Table 4.12 where we display the mean, standard de-

[11] The share is computed for each individual, leaving his father out when computing the share.

viation, 10th and 90th percentile of the share of fathers graduated across different fields and the residuals from the regression of the share of fathers graduated in different fields on school fixed effects and school fixed effects interacted with the shares of maximum education of the fathers of peer students (colum *residualized*). In the column *plain*, we see that on average a mathematical high school student in her last year of school has 2% of fathers peers' graduated in humanities, with a standard deviation of 0.026. To obtain a sense of the variability in the share of fathers peers' graduated in humanities across difference school-years cells, we show the 10th and 90th percentile of the share of fathers peers' humanities graduates[12]. The least *treated* students, meaning individuals exposed to the 10th percentile of the share of peers with fathers graduated in humanities, have a 0% share. That is, in their school-year cell, nobody is the son of a father graduated in humanities. The very *treated* individuals are students exposed to the 90th percentile of the share of fathers peers' graduated in humanities and have a 6.3% share of peer students with fathers graduated in humanities in their school-year cell. We additionally display the same moments for the other relevant fields of study: business, engineering, and natural sciences. In general, the share of fathers peers' graduates across the different disciplines can be thought as the different probability to meet a peer student in the school that has higher abilities in the field of his father's university subject and represents for the other students a source of accumulation of field-specific human capital through peer interactions and learning spillovers. Our identification strategy relies on exploiting different exposures of students to the accumulation of specific skills that arise from the human capital of fathers peers' students. In the main regression 4.4 we control for the overall level of fathers with a university degree. This makes sure that we do not use variation in the share of fathers peers' with humanities degrees that only come from school-years cells having a higher overall number of fathers with a university degree. For this reason, Table 4.12 in column *residualized* shows the moments of the share of fathers peers' graduated across different disciplines after controlling for school fixed effects and school fixed effects interacted with the shares of maximum education of the fathers of peer students. The standard deviation is around 0.016 and the difference between the least and the most treated individual is about a 3.5% difference in the share of peers with fathers graduated in humanities.

[12]For privacy policies when using the data we cannot show percentiles of distributions, but show instead averages around these percentiles.

Table 4.12: Variability in treatment

	Plain				Residualized			
	Mean	St. Dev	mean 9-11 pc	mean 89-91 pc	Mean	St. Dev	mean 9-11 pc	mean 89-91 pc
Humanities	0.025	.030	0	0.063	0	.016	-0.017	0.017
Business	0.029	0.033	0	0.072	0	.016	-0.017	0.018
Engineering	.039	0.043	0	0.11	0	.018	-0.021	0.020
Natural sciences	0.025	0.028	0	0.065	0	0.015	-0.016	0.017

Notes: The table reports mean, St. Dev, mean 9-11 percentile, mean 89-91 percentile of the share of peers' fathers graduated in humanities and residuals of a regression of the share of peers' fathers graduated in humanities on school f.e. and school f.e. interacted with the education share of peers parents (Residualized), for Mathetimatical HS students in their last year belonging to the cohort 1997-2004.

4.8.2 Father field of graduation and their son performance

Given that our empirical strategy relies on using human capital of fathers peer students' as an instrument for the communication skills of an individual, in this section we provide evidence on the intergenerational transmission of human capital and skills between fathers and sons. Specifically, in Table 4.13 we show the average grade in math and danish, high school gpa and the probability of becoming an entrepreneur of high school students, for the different disciplines in which their fathers obtained a university degrees. The sample is made of all high school students graduated in a mathematical high school between 1997 and 2004, who have fathers with a university degree. The averages in the table are computed with respect to a reference made of the averages of sons with fathers graduated in other disciplines: education, social science, information, agriculture, welfare, service and unknown. In column (1) we show that the average grade in danish for students whose father is graduated in humanities is significantly higher than the average of students who have fathers graduated in other disciplines (almost 7% higher than the reference). In particular, the grade is higher than the one obtained by sons of graduates in business, natural sciences and engineering who all have lower than reference average grades. The same pattern applies when we consider oral and written examinations separately, with a greater positive difference between the grades in oral than written exams (column (2) and (3)). When we move to the grade in math (column (4)) we see that sons of humanity graduates have a higher average grade also in math, but the positive difference with respect to the reference category is smaller, while significant (2.8% higher grade). Sons of business graduates do not have a significant different average grade compared to the reference category, while the sons of graduates in natural sciences have a significant higher grade in math (6.7% higher). Somehow surprisingly, sons of engineers have lower than average grade in math,

and also lower than the average math grade of sons of humanity graduates.
In Column (5) we display the average gpa in the last year of high school by father field of graduation. Students with fathers graduated in humanities are those with the highest gpa. Finally, in column (6) we show the average share of entrepreneurs by the field of graduation of the father. The son of humanity graduates have on average a 1,2 percentage point lower share of entrepreneurs than the reference category, but the difference is not significant. On the other hand, as expected, sons of business graduates have a 2.5 percentage points higher share of entrepreneurs compared to the baseline (which is more than the half of the average share of entrepreneurs in the sample).
In light of the evidence in Table 4.13, we showed that indeed the son of a graduate in humanities has better communication skills, as proxied by the grade in the oral danish exams, than sons of graduates in other disciplines. This evidence is suggestive of the fact that human capital is transmitted across generations and supports our empirical approach. Second, sons of humanity graduates do not become more frequently entrepreneurs, as is the case instead for sons of business graduates. This is important as it implies that there are no other factors pushing sons of humanity graduates to select into entrepreneurship. Finally, the literature on intergenerational skill transmission usually thinks of human capital being transmitted between father and sons as a bundle of skills. While it is true that sons of humanity graduates have higher gpa on average, we know from the previous sections that high school gpa is negatively associated with the probability of becoming an entrepreneur.

4.9 Second stage

In this section we analyze the effect of being exposed to a higher share of peers whose father is graduated in humanities on the probability to become an entrepreneur. If the two identifying assumptions hold, then equation 4.4 captures the casual effect of increasing communication skills of mathematical high school students on the probability to start a business and become entrepreneurs.

In Table 4.14, we report the estimated regression coefficients of equation 4.4. We display the coefficients estimated on subsamples of different cohorts: in (1) cohorts from 1997 to 2004, in (2) from 1997 to 2003, in (3) from 1997 to 2002 and in (4) from 1997 to 2001. The reason for this goes back to our initial findings that the majority of individuals open a business relatively late in life, implying that later cohorts might not have had the time yet to become entrepreneurs. When we con-

Table 4.13: Father field of graduation and son high school grades

	(1) Danish	(2) Danish Oral	(3) Danish Written	(4) Math	(5) log-gpa	(6) Entrepreneur
Humanities	0.643***	0.641***	0.577***	0.267***	0.0437***	-0.0119
	(0.111)	(0.132)	(0.117)	(0.146)	(0.0134)	(0.00728)
Business	-0.169*	-0.116	-0.192**	-0.0971	-0.0141	0.0248**
	(0.0901)	(0.113)	(0.0890)	(0.145)	(0.0126)	(0.00931)
Natural Science	-0.246***	-0.305**	-0.185**	0.635***	0.00958	0.00484
	(0.0902)	(0.114)	(0.0900)	(0.143)	(0.0123)	(0.00851)
Engineering	-0.487***	-0.475***	-0.478***	-0.245**	-0.0717***	0.0172**
	(0.0776)	(0.0976)	(0.0792)	(0.122)	(0.0116)	(0.00741)
_cons	9.307***	9.565***	9.056***	9.450***	4.355***	0.0431***
	(0.0435)	(0.0534)	(0.0453)	(0.0692)	(0.00583)	(0.00376)
N	7050	7050	7050	7050	7050	7050

Notes: $^{*}(p<0.1),^{**}(p<0.05),^{***}(p<0.01)$. This table reports OLS coefficients of the regression of the average grade in the final year of high school (for student of Mathematical HS) in (1) Danish, (2)Oral Danish, (3) Written Danish, (4) Math, (5) probability of becoming an entrepreneur and (6) log-gpa on a dummy for the father of an individual being graduated in humanities, business, Natural Science, Engineering. The coefficients display the difference in the variable of interest of the mean of every field with respect to the rest of the group of fields that are: education, social sciences, information, agriculture, welfare, service, unknown. An individual is considered attending Mathematical HS if she attended Mathematical HS and took level A math (80% of the sample of mathematical HS students). The sample is the universe of all Mathematical HS students having a father that has a university education. Standard errors are clustered at school year level. Cluster #: 1053 all specification

sider the cohorts 1997 to 2004 in column (1), the effect of increasing the share of fathers peers' with a university diploma in humanities from 0% to 100% increases the probability to become an entrepreneur by 7.4 percentage points, but is not significantly different from zero. When we start to remove later cohorts- columns (2) to (4)- we see that the effect increases up to 31 percentage points (column (4)). If we assume that the estimated effect for the cohorts 1997-2001 is close to the true one, if we could have followed individuals of these cohorts over their entire life-cycle, this implies that individuals being exposed to peers with all fathers graduated in humanities compared to none, increases the probability of ever becoming an entrepreneur by 31.5 percentage points. One has to consider, however, that the difference between the least and the most treated individual in our dataset -the difference between the 90th and 10th percentile - is 3.5 percentage points. Thus the difference in the probability of becoming an entrepreneur between an individual who is the least and the most treated is 1.1 percentage points. To put this into context, this corresponds to 20% of the overall share of entrepreneurs in the economy for the 1997-2001 cohort. This is an economically significant effect.

Table 4.14: Second stage regression

	1997-2004	1997-2003	1997-2002	1997-2001
	(1)	(2)	(3)	(4)
	Entrepreneur	Entrepreneur	Entrepreneur	Entrepreneur
1) Share_humanities	0.0737	0.165**	0.249**	0.316**
	(0.0649)	(0.0771)	(0.102)	(0.148)
2) Share_business	0.154**	0.190**	0.0184	0.161
	(0.0717)	(0.0872)	(0.132)	(0.157)
3) Share_natural_sciences	-0.0606	-0.0772	-0.0586	-0.0928
	(0.0727)	(0.0858)	(0.115)	(0.170)
4) Share_engineering	0.0509	0.0428	0.0728	0.223
	(0.0647)	(0.0764)	(0.107)	(0.151)
School f.e.	✓	✓	✓	✓
School f.e.×Share_HS+9	✓	✓	✓	✓
School f.e.×Share_HS+12	✓	✓	✓	✓
School f.e.×Share_HS+14	✓	✓	✓	✓
School f.e.×Share_HS+15	✓	✓	✓	✓
School f.e.×Share_HS+17	✓	✓	✓	✓
School f.e.×Share_HS+20	✓	✓	✓	✓
Father_year_of_education×education_area	✓	✓	✓	✓
High school municipality f.e.	✓	✓	✓	✓
N	29642	26291	23051	19668

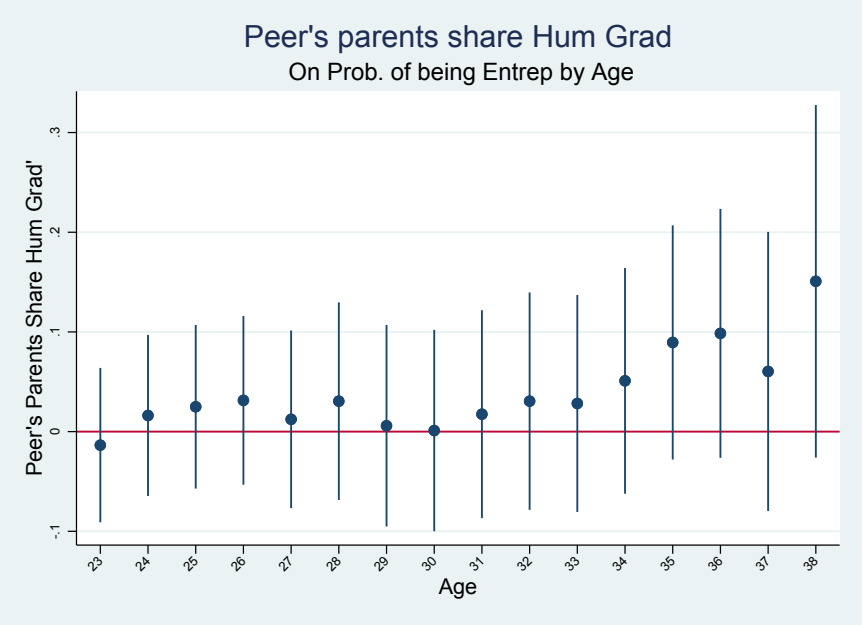
Notes: $^{*}(p < 0.1),^{**}(p < 0.05),^{***}(p < 0.01)$. This table reports the OLS coefficients of the regressing the probability of ever becoming an entrepreneur on the share of peers whose father is graduated in humanities, controlling for school f.e., school f.e. interacted with share of education of peers' fathers, father year of education interacted with field of education and high school municipality f.e. In column (1) the sample contains cohort from 1997 to 2004, column (2) cohorts from 1997 to 2003, column (3) cohorts 1997 to 2003, and column (4) cohorts 1997 to 2002. In the second, third and fourth row we report the same OLS coefficients where instead of using the share of peers' fathers with a degree in humanities we use those with a degree in business, natural sciences and engineering respectively. The sample is the universe of male Danish Mathematical high school students in their last HS school year. *Entrepreneur* is 1 if an individual has ever been an entrepreneur in her life. Standard errors are clustered at year-school level and are: # (1) 1,109, (2) 969, (3) 829, (4) 691

In the second row of Table 4.14, we report the regression coefficients of equation 4.4 in which we use the share of fathers peers' with a degree in business- instead of humanities- for different cohorts. We do this as an indirect test of our identification strategy, because as shown by Mertz et al. (2023) sons of business graduates are likelier to start a business. This also helps us to have a benchmark against which to compare the estimated coefficients for humanities. In column (1) we see that increasing the share of fathers peers' with a university diploma in business from 0% to 100% increases the probability to become an entrepreneur by 16 percentage points, which becomes 19 percentage points for cohorts 1997-2003 (column (2)). Comparing for the same cohorts the effect of increasing the share of fathers peers' with humanities or business degrees has similar effects in magnitude.

This means that improving the communication skills of students who are good at math or providing more early exposure of students to an entrepreneurial environment has a quantitative comparable effect on spurring new business creation.

Finally, in the second and third row of Table 4.14 we show the same results but changing fathers' university degree to natural sciences and engineering respectively. We see that being expose to a higher share of students with fathers graduated in natural sciences has a negative effect, even if statistically not different from zero, on the probability of becoming an entrepreneur. This aligns with our findings in section 7.2 that fathers graduated in natural sciences on average transmit higher mathematical skills to their sons, which alone negatively relate to self-selection into entrepreneurship. For fathers graduated in engineering the effect on the probability of ever becoming an entrepreneur is positive, but again not statistically significant. In Figures 4.4 and 4.5, we present the coefficient β_3 from regression 4.4 for the population of high school mathematical graduates from the 1997 to 2004 cohorts for a specfication where the share of father peers graduates in humanities and business is interacted with the age of the individual. As depicted in the figures, the effect of the share of father peers graduates increases with age. Noteworthy the effect of exposure to humanities starts increasing after the age of 34 when the probability of opening a business is at its maximum (see Figure 4.1). The effect for exposure to business starts increasing earlier, at 30 years old. While the effect for a one hundred percent variation in humanities fathers peers graduates for an individual at 38 years old is nearly 20 percentage points, which is two-thirds of the effect for exposure to business father graduates (see Figure 4.5), the estimates for humanities are imprecise. In future analyses, we plan to include the mother peers' human capital field of graduation to enhance the precision of humanities effect estimates.
Taken together, this evidence shows that mathematical high school students in their last year of school, who were exposed to a higher share of peers with fathers graduated in humanities - for a given school and given overall number of fathers with university diploma - have a significantly higher probability of becoming entrepreneurs compared to the rest of the population. The effect is sizable, being around 20% of the share of entrepreneurs in the economy for the 1997-2021 cohorts. The effect is also comparable, in terms of magnitudes, to exposing students to more entrepreneurial environments as measured by the share of fathers peer students' graduated in business.

Figure 4.4: Share of humanities graduates parents peers and prob. to become entrepreneur by age

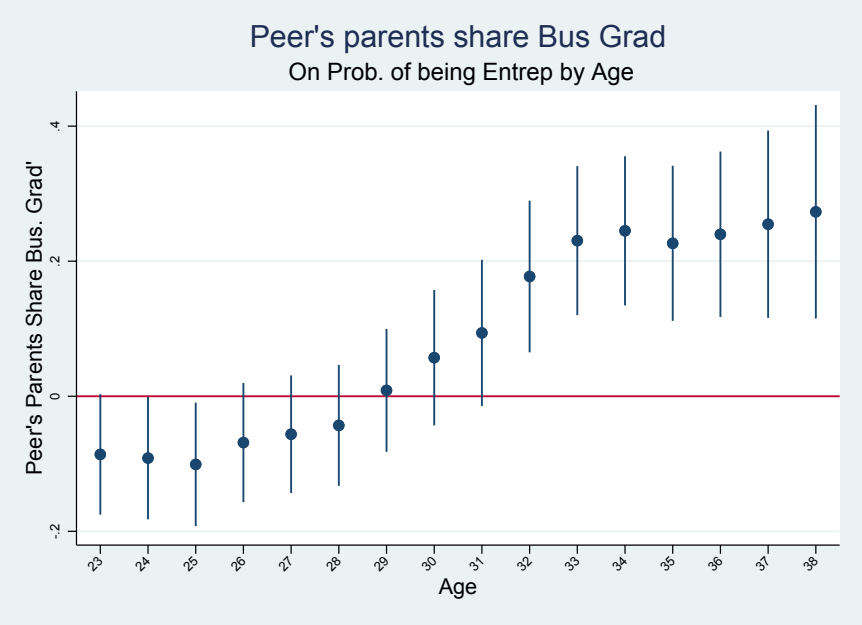


Notes— The figure displays the coefficient β_3 from regression 4.4, for the specification where $Hum_{J-i,i}$ is interacted with the age of the individual. Coefficients report the variation in the probability of being an entrepreneur at every age for a one hundred percent variation in the share of peers' fathers who are humanities graduates, for given school and proportion of H.S. +5 fathers. Population: H.S. math graduates between 1997 and 2004. Period of observation: 1997 to 2019.

4.10 Conclusions

In this paper we use Danish administrative data to provide new findings on the role of skill multidimensionality in explaining self-selection into entrepreneurship. We use detailed high school grades obtained by students in math and danish to measure analytical and communication skills. Observational evidence shows that individuals with specialized skills earn higher wages on the labor market, but are less likely to become entrepreneurs. For high talented math students in high school, the probability of starting a business is increasing in their communication skills. These students are also the ones owning the most profitable and successful businesses if they become entrepreneurs. To casually estimate the effect of improving communication skills of individuals with good analytical skills, we propose an identification strategy that exploits information on parents' human capital. Our findings show that teaching better communication skill to students who are very good at math in

Figure 4.5: Share of business graduates parents peers and prob. to become entrepreneur by age



Notes— The figure displays the coefficient β_3 from regression 4.4, for the specification where $Bus_{J-i,i}$ is interacted with the age of the individual. Coefficients report the variation in the probability of being an entrepreneur at every age for a one hundred percent variation in the share of peers father who are business graduates, for given school and given share of graduated father. Population: H.S. math graduates between 1997 and 2004. Period of observation: 1997 to 2019.

high school spurs business creation and that the effects are sizable. Our findings contribute to growing evidence on the role of human capital for entrepreneurial outcomes and inform the policy debate on the importance of education and training programs to incentivize the birth of successful entrepreneurs.

Bibliography

Abrahams, Fred (2016). *Modern Albania: from dictatorship to democracy in Europe*. NYU Press, URL https://doi.org/10.18574/nyu/9780814705117.001.0001. OCLC: 967257139.

Acemoglu, Daron (1999). "Changes in unemployment and wage inequality: An alternative theory and some evidence." *American economic review*, 89(5), 1259–1278.

Acemoglu, Daron and David Autor (2011). "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics*, vol. 4, pp. 1043–1171. Elsevier.

Adema, Joop, Cevat Giray Aksoy, and Panu Poutvaara (2022). "Mobile Internet Access and the Desire to Emigrate." *CESifo Working Papers*.

Adsera, Alicia and Ana M. Ferrer (2015). "The Effect of Linguistic Proximity on the Occupational Assimilation of Immigrant Men in Canada." *IZA Discussion Paper*.

Adserà, Alícia and Mariola Pytliková (2015). "The Role of Language in Shaping International Migration." *The Economic Journal*, 125(586), F49–F81.

Allman, Elisabeth S., Catherine Matias, and John A. Rhodes (2009). "Identifiability of Parameters in Latent Structure Models with Many Observed Variables." *Annals of Statistics*, 37(6A), 3099–30132.

Anelli, Massimo, Gaetano Basso, Giuseppe Ippedico, and Giovanni Peri (2023). "Emigration and Entrepreneurial Drain." *American Economic Journal: Applied Economics*, 15(2), 218–252.

Arcidiacono, Peter, Esteban Aucejo, Arnaud Maurel, and Tyler Ransom (2016). "College Attrition and the Dynamics of Information Revelation." Working Paper

22325, National Bureau of Economic Research, URL http://www.nber.org/papers/w22325.

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano (2020). "Ex Ante Returns and Occupational Choice." *Journal of Political Economy*, 128(12), 4475–4522.

Arcidiacono, Peter and John Bailey Jones (2003). "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm." *Econometrica*, 71(3), 933–946.

Argan, Damiano and Robert J. Gary-Bobo (2021). "Les diplômes français se sont-ils dévalorisés?" *Revue Economique, forthcoming*.

Argan, Damiano, Leonardo Indraccolo, and Jacek Piosik (2023). "Teach the Nerds to Make a Pitch: Multidimensional Skills and Selection into Entrepreneurship."

Ashworth, Jared, V Joseph Hotz, Arnaud Maurel, and Tyler Ransom (2021). "Changes across cohorts in wage returns to schooling and early work experiences." *Journal of labor economics*, 39(4), 931–964.

Autor, David, Claudia Goldin, and Lawrence F. Katz (2020). "Extending the Race between Education and Technology." *AEA Papers and Proceedings*, 110, 347–51.

Autor, David H, Lawrence F Katz, and Melissa S Kearney (2008). "Trends in US wage inequality: Revising the revisionists." *The Review of economics and statistics*, 90(2), 300–323.

Azoulay, Pierre, Benjamin F. Jones, J. Daniel Kim, and Javier Miranda (2020). "Age and High-Growth Entrepreneurship." *American Economic Review: Insights*, 2(1), 65–82.

Beaudry, Paul, David A Green, and Benjamin M Sand (2014). "The declining fortunes of the young since 2000." *American Economic Review*, 104(5), 381–86.

Beaudry, Paul, David A Green, and Benjamin M Sand (2016). "The great reversal in the demand for skill and cognitive tasks." *Journal of Labor Economics*, 34(S1), S199–S247.

Beffy, Magali, Denis Fougère, and Arnaud Maurel (2012). "Choosing the Field of Study in Postsecondary Education: Do Expected Earnings Matter?" *The Review of Economics and Statistics*, 94(1), 334–347.

Belot, Michèle and Sjef Ederveen (2012). "Cultural barriers in migration between OECD countries." *Journal of Population Economics*, 25(3), 1077–1105.

Belot, Michèle and Timothy Hatton (2012). "Immigrant Selection in the OECD." *The Scandinavian Journal of Economics*, 114(4), 1105–1128.

Belzil, Christian and Jorgen Hansen (2020). "Reconciling changes in wage inequality with changes in college selectivity, using a behavioral model." Institut Polytechnique de Paris.

Berman, Eli, Kevin Lang, and Erez Siniver (2003). "Language-skill complementarity: returns to immigrant language acquisition." *Labour Economics*, 10(3), 265–290.

Blanchflower, David G and Andrew J Oswald (1998). "What Makes an Entrepreneur?" *Journal of Labor Economics*.

Blundell, Richard, David A. Green, and Wenchao Jin (2022). "The U.K. as a Technological Follower: Higher Education Expansion and the College Wage Premium." *Review of Economic Studies*, 89, 142–180.

Bonhomme, Stéphane, Koen Jochmans, and Jean-Marc Robin (2016). "Estimating multivariate latent-structure models." *Annals of Statistics*, 44(2), 540–563.

Borjas, George J (1987). "Self-Selection and the Earnings of Immigrants." *The American Economic Review*, 77, 24.

Bouveyron, Charles, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery (2019). *Model-based clustering and classification for data science: with applications in R*, vol. 50. Cambridge University Press.

Braga, Michela (2007). "Dreaming Another Life. The Role of Foreign Media in Migration Decision. Evidence from Albania." *World Bank*, p. 41.

Bursztyn, Leonardo and Davide Cantoni (2016). "A Tear in the Iron Curtain: The Impact of Western Television on Consumption Behavior." *Review of Economics and Statistics*, 98(1), 25–41.

Bütikofer, Aline and Giovanni Peri (2021). "How Cognitive Ability and Personality Traits Affect Geographic Mobility." *Journal of Labor Economics*, 39(2), 559–595.

Card, David and Thomas Lemieux (2001). "Can falling supply explain the rising return to college for younger men? A cohort-based analysis." *The quarterly journal of economics*, 116(2), 705–746.

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman (2003). "2001 Lawrence R. Klein Lecture Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice*." *International Economic Review*, 44(2), 361–422.

Carneiro, Pedro, James J. Heckman, and Edward J. Vytlacil (2011). "Estimating Marginal Returns to Education." *American Economic Review*, 101(6), 2754–81.

Carneiro, Pedro and Sokbae Lee (2011). "Trends in quality-adjusted skill premia in the United States, 1960-2000." *American Economic Review*, 101(6), 2309–49.

Cassagneau-Francis, Oliver (2021). "The Returns to Higher Education by Cognitive and Non-cognitive Abilities." Tech. rep., UCL, London.

Cassagneau-Francis, Oliver, Robert Gary-Bobo, Julie Pernaudet, and Jean-Marc Robin (2021). "A Non-parametric Finite Mixture Approach to Difference-in-Difference Estimation, with an Application to On-the-job Trainig and Wages." Tech. rep., Sciences Po, Paris.

Chiswick, Barry R (1995). "The Endogeneity between Language and Earnings: International Analyses." *Journal of Labor Economics*.

Choi, Joonkyu, Nathan Goldschlag, John Haltiwanger, and J Daniel Kim (2019). "Founding teams and startup performance." *Available at SSRN 3481850*.

Chong, Alberto and Eliana La Ferrara (2009). "Television and Divorce: Evidence from Brazilian *Novelas*." *Journal of the European Economic Association*, 7(2-3), 458–468.

Corblet, Pauline (2022). "The Decreasing Returns to Experience for Higher Education Graduates in France: An Occupational Analysis." Tech. rep., Sciences Po, Paris.

Cunha, Flavio, James Heckman, and Salvador Navarro (2005). "Separating uncertainty from heterogeneity in life cycle earnings." *Oxford Economic Papers*, 57(2), 191–261.

Cunha, Flavio and James J. Heckman (2007). "Identifying and Estimating the Distributions of Ex Post and Ex Ante Returns to Schooling." *Labour Economics*, 14(6), 870–893. Education and Risk.

Docquier, Frédéric and Abdeslam Marfouk (2006). "International Migration by Education Attainment, 1990-2000." In *International migration, remittances, and the brain drain*, edited by Çaglar Özden and Maurice W. Schiff, Trade and development series, chap. 5, pp. 151–200. World Bank : Palgrave Macmillan, Washington, DC.

Docquier, Frédéric and Hillel Rapoport (2012). "Globalization, Brain Drain, and Development." *Journal of Economic Literature*, 50(3), 681–730.

Dorfles, Piero and Giovanna Gatteschi (1991). *Guardando all'Italia : influenza delle TV e delle radio italiane sull'esodo degli albanesi*. No. 1 in Instant research / VQPT-SO, 1 ed., Roma : RAI radiotelevisione italiana.

Durante, Ruben, Paolo Pinotti, and Andrea Tesei (2019). "The Political Legacy of Entertainment TV." *American Economic Review*, 109(7), 2497–2530.

Dustmann, Christian and Costas Meghir (2005). "Wages, experience and seniority." *The Review of Economic Studies*, 72(1), 77–108.

Emmons, William R, Ana HernÃ Kent, Lowell Ricketts, et al. (2019). "Is college still worth it? The new calculus of falling returns." *Federal Reserve Bank of St Louis Review*, 101(4), 297–329.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya (2011). "Media and Political Persuasion: Evidence from Russia." *American Economic Review*, 101(7), 3253–3285.

Etgeton, Stefan (2018). "Cluster-SE." *Github*.

Farré, Lídia and Francesco Fasani (2013). "Media exposure and internal migration — Evidence from Indonesia." *Journal of Development Economics*, 102, 48–61.

Ferrara, Eliana La, Alberto Chong, and Suzanne Duryea (2012). "Soap Operas and Fertility: Evidence from Brazil." *American Economic Journal: Applied Economics*, 4(4), 1–31.

Fevziu, Blendi, Robert Elsie, Majlinda Nishku, and Blendi Fevziu (2018). *Enver Hoxha: the iron fist of Albania*. Bloomsbury Publishing, URL https://doi.org/10.5040/9781350986268?locatt=label:secondary_bloomsburyCollections. OCLC: 1128170332.

Galanxhi, Emira, Elena Misja, Desareta Lameborshi, Mathias Lerch, Philippe Wanner, and Janine Dahinden (2004). *Migration in Albania: population and housing census 2001*. No. 18 in 2001 population and housing census, INSTAT, Tirane.

Gary-Bobo, Robert J, Marion Goussé, and Jean-Marc Robin (2016). "Grade retention and unobserved heterogeneity." *Quantitative Economics*, 7(3), 781–820.

Gentzkow, Matthew and Jesse M. Shapiro (2008). "Preschool Television Viewing and Adolescent Test Scores: Historical Evidence from the Coleman Study *." *Quarterly Journal of Economics*, 123(1), 279–323.

Goldin, Claudia and Lawrence F Katz (2008). *The race between education and technology*. harvard university press.

Guiso, Luigi, Luigi Pistaferri, and Fabiano Schivardi (2021a). "Learning Entrepreneurship from Other Entrepreneurs?" *Journal of Labor Economics*.

Guiso, Luigi, Luigi Pistaferri, and Fabiano Schivardi (2021b). "Learning entrepreneurship from other entrepreneurs?" *Journal of labor economics*, 39(1), 135–191.

Guvenen, Fatih (2007). "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *American Economic Review*, 97(3), 687–712.

Guvenen, Fatih, Fatih Karahan, Serdar Ozkan, and Jae Song (2021). "What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics?" *Econometrica*, 89(5), 2303–2339.

Gëdeshi, Ilir and Russell King (2019). "The Albanian scientific diaspora: can the brain drain be reversed?" *Migration and Development*, 10(1), 19–41.

Gërmenji, Etleva and Lindita Milo (2011). "Migration of the skilled from Albania: brain drain or brain gain?" *Journal of Balkan and Near Eastern Studies*, 13(3), 339–356.

Hall, Peter and Xiao-Hua Zhou (2003). "Nonparametric estimation of component distributions in a multivariate mixture." *The Annals of Statistics*, 31(1), 201 – 224.

Heckman, James J., John Eric Humphries, and Gregory Veramendi (2018). "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking." *Journal of Political Economy*, 126(S1), S197–S246.

Heckman, James J. and Brain Singer (1984). "A Method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica*, 52, 271–320.

Heckman, James J and Edward Vytlacil (2005). "Structural equations, treatment effects, and econometric policy evaluation 1." *Econometrica*, 73(3), 669–738.

House, Christopher, Christian Proebsting, and Linda Tesar (2018). "Quantifying the Benefits of Labor Mobility in a Currency Union." Tech. Rep. w25347, National Bureau of Economic Research, Cambridge, MA, URL http://www.nber.org/papers/w25347.pdf.

Huggett, Mark, Gustavo Ventura, and Amir Yaron (2011). "Sources of Lifetime Inequality." *American Economic Review*, 101(7), 2923–2954.

Ichino, Andrea, Aldo Rustichini, and Giulio Zanella (2022). "College Education, Intelligence and Disadvantage: Policy Lessons from the UK, 1960-2004."

Inoue, Atsushi and Gary Solon (2010). "Two-Sample Instrumental Variables Estimators." *The Review of Economics and Statistics*, 92(3), 557–561.

Jensen, Robert and Emily Oster (2009). "The Power of TV: Cable Television and Women's Status in India *." *Quarterly Journal of Economics*, 124(3), 1057–1094.

Jeong, Hyeok, Yong Kim, and Iourii Manovskii (2015). "The price of experience." *American Economic Review*, 105(2), 784–815.

Kambourov, Gueorgui and Iourii Manovskii (2009). "Occupational specificity of human capital." *International Economic Review*, 50(1), 63–115.

Kasahara, Hiroyuki and Katsumi Shimotsu (2009). "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices." *Econometrica*, 77, 135–175.

Katz, Lawrence F and Kevin M Murphy (1992). "Changes in relative wages, 1963–1987: supply and demand factors." *The quarterly journal of economics*, 107(1), 35–78.

Kearney, Melissa S. and Phillip B. Levine (2015). "Media Influences on Social Outcomes: The Impact of MTV's *16 and Pregnant* on Teen Childbearing." *American Economic Review*, 105(12), 3597–3632.

Kearney, Melissa S. and Phillip B. Levine (2019). "Early Childhood Education by Television: Lessons from *Sesame Street*." *American Economic Journal: Applied Economics*, 11(1), 318–350.

La Ferrara, Eliana (2016). "Mass Media and Social Change: Can we Use Television to Fight Poverty?" *Journal of the European Economic Association*, 14(4), 791–827.

Lang, Julia (2022). "Employment effects of language training for unemployed immigrants." *Journal of Population Economics*, 35(2), 719–754.

Lazear, Edward P (2004a). "Balanced Skills and Entrepreneurship." *American Economic Review*, 94(2), 208–211.

Lazear, Edward P (2004b). "Balanced skills and entrepreneurship." *American Economic Review*, 94(2), 208–211.

Lazear, Edward P (2005). "Entrepreneurship." *Journal of Labor Economics*.

Levine, Ross and Yona Rubinstein (2017a). "Smart and Illicit: Who Becomes an Entrepreneur and Do They Earn More?*." *The Quarterly Journal of Economics*, 132(2), 963–1018.

Levine, Ross and Yona Rubinstein (2017b). "Smart and illicit: who becomes an entrepreneur and do they earn more?" *The Quarterly Journal of Economics*, 132(2), 963–1018.

Liang, James, Hui Wang, and Edward P Lazear (2018). "Demographics and Entrepreneurship." *Journal of Political Economy*.

Lochmann, Alexia, Hillel Rapoport, and Biagio Speciale (2019). "The effect of language training on immigrants' economic integration: Empirical evidence from France." *European Economic Review*, 113, 265–296.

Lucas, Robert E (1978). "On the Size Distribution of Business Firms." *The Bell Journal of Economics*.

Magnac, Thierry, Nicolas Pistolesi, and Sébastien Roux (2018). "Post-Schooling Human Capital Investments and the Life Cycle of Earnings." *Journal of Political Economy*, 126(3), 1219–1249.

Magnac, Thierry and David Thesmar (2002). "Identifying Dynamic Discrete Decision Processes." *Econometrica*, 70(2), 801–816.

Mai, Nick (2004). "'Looking for a More Modern Life...': the Role of Italian Television in the Albanian Migration to Italy." *Westminster Papers in Communication and Culture*, 1(1), 3.

McKenzie, David, John Gibson, and Steven Stillman (2013). "A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad?" *Journal of Development Economics*, 102, 116–127.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. Wiley series in probability and statistics: Applied probability and statistics, Wiley, URL https://books.google.fr/books?id=7M5vK8OpXZ4C.

Mertz, Mikkel Baggesgaard, Maddalena Ronchi, and Viola Salvestrini (2023). "Early exposure to entrepreneurs, gender equality, and talent allocation in entrepreneurship."

Michelacci, Claudio and Fabiano Schivardi (2020a). "Are they all like Bill, Mark, and Steve? The education premium for entrepreneurs." *Labour Economics*, 67, 101933.

Michelacci, Claudio and Fabiano Schivardi (2020b). "Are they all like Bill, Mark, and Steve? The education premium for entrepreneurs." *Labour Economics*, 67, 101933.

Mundell, Robert (1961). "A Theory of Optimum Currency Areas." *The American Economic Review*, 51(4), 657–665.

Olken, Benjamin A. (2009). "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages." *American Economic Journal: Applied Economics*, 1(4), 1–33.

Pacini, David and Frank Windmeijer (2016). "Robust inference for the Two-Sample 2SLS estimator." *Economics Letters*, 146, 50–54.

Pesando, Luca Maria, Valentina Rotondi, Manuela Stranges, Ridhi Kashyap, and Francesco C. Billari (2021). "The Internetization of International Migration." *Population and Development Review*, 47(1), 79–111.

Queiró, Francisco (2022). "Entrepreneurial human capital and firm dynamics." *The Review of Economic Studies*, 89(4), 2061–2100.

Riley, Shawn, Stephen DeGloria, and Robert Elliot (1999). "A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity." *Intermountain Journal of Sciences*, 5(1-4).

Sarvimäki, Matti and Kari Hämäläinen (2016). "Integrating Immigrants: The Impact of Restructuring Active Labor Market Programs." *Journal of Labor Economics*, 34(2), 479–508.

Schmid, Lukas (forthcoming). "The Impact of Host Language Proficiency on Migrants' Employment Outcomes." *American Economic Review: Insights*.

Shrestha, Slesh A. (2017). "No Man Left Behind: Effects of Emigration Prospects on Educational and Labour Outcomes of Non-migrants." *The Economic Journal*, 127(600), 495–521.

Silva, Olmo (2007). "The Jack-of-All-Trades entrepreneur: Innate talent or acquired skill?" *Economics Letters*, 97(2), 118–123.

Valletta, Robert G (2018). "Recent flattening in the higher education wage premium: Polarization, skill downgrading, or both?" In *Education, skills, and technical change: Implications for future US GDP growth*, pp. 313–342. University of Chicago Press.

Verdugo, Gregory (2014). "The great compression of the French wage structure, 1969–2008." *Labour Economics*, 28, 131–144.

Wagner, J. (2003). "Testing Lazear's jack-of-all-trades view of entrepreneurship with German micro data." *Applied Economics Letters*, 10(11), 687–689.

Wagner, Joachim (2006). "Are nascent entrepreneurs 'Jacks-of-all-trades'? A test of Lazear's theory of entrepreneurship with German data." *Applied Economics*, 38(20), 2415–2419.

Appendix A

First Appendix

A.1 Italian TV Shows Watched by Albanian Migrants

In this section we report the distribution of TV shows watched by Albanians migrants. In 1991, Dorfles and Gatteschi (1991) interviewed 311 Italian speaking Albanian migrants just arrived in Italy, 301 declared watching Italian television back home in Albania. They were asked to list all Italian shows they usually watched in Albania. Table A.1 (extracted from Dorfles and Gatteschi (1991)) reports the results. We report a brief description and a Wikipedia link for the TV shows that count at least 4% of answer: they are all entertainment shows. *TG1*, the main Italian news show, appears in less than 3% of answers.

Domenica In: "Domenica in is an entertainment Italian TV show on air on Rai 1 since 1976.." (Wikipedia:https://it.wikipedia.org/wiki/Domenica_in)

Fantastico: "Fantastico was an Italian TV variety show broadcast saturday prime time on Rai 1 from 1979 to 1980 and from 1981 to 1992.."(Wikipedia: https://it.wikipedia.org/wiki/Fantastico_(programma_televisivo)

Piacere RaiUno:"During the show there were prank calls, dance, music and interview to popular TV characters. There were also some time dedicated to information about issues of different Italian city" (Wikipedia: https://it.wikipedia.org/wiki/Piacere_Raiuno)

La Domenica Sportiva: "La Domenica Sportiva is the oldest sport show of Italian television." (Wikipedia:https://it.wikipedia.org/wiki/La_Domenica_Sportiva)

Crème Caramel: variety and Vaudeville Tv show (Wikipedia: https://

it.wikipedia.org/wiki/Cr%C3%A8me_Caramel_(programma_televisivo)

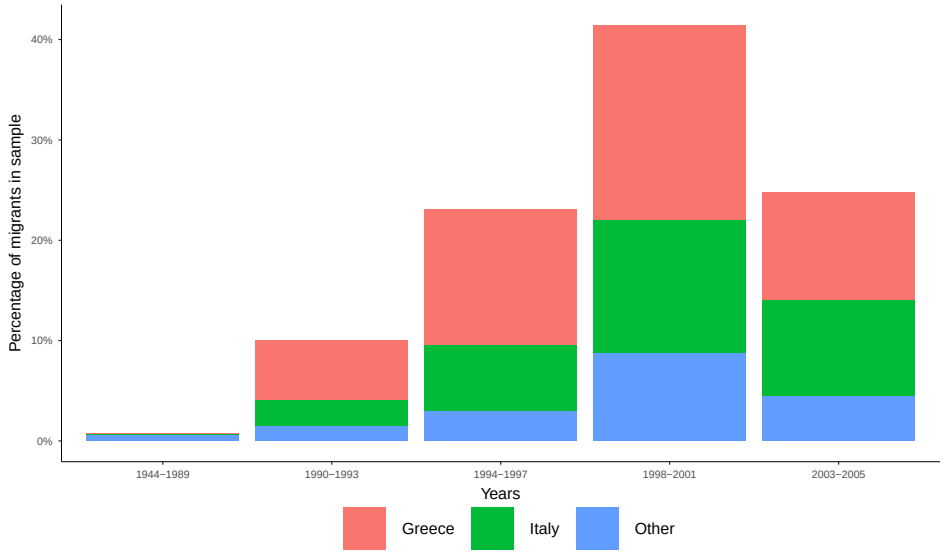
Quark: TV show to popularize science (Wikipedia: https://it.wikipedia.org/wiki/Quark_(programma_televisivo))

Sanremo: Broadcasted music festival (Wikipedia: https://it.wikipedia.org/wiki/Festival_di_Sanremo)

A.2 Emigration Patterns of Albanians 1990-2005

Figure A.1 reports the distribution of migrants over time and over destination country in the sample. For each 4 years bracket, we compute the share of the migrants in the sample and their destination country choice. We can see that the share of migrants prior to 1990 (end of the regime) is almost null, and that it increasing until 1998-2001, when it peaks before decreasing afterwards.

Figure A.1: Emigration Patterns of Albanians 1990-2005. Source: *Siblings sample* from 2005 Living Standard Measurement Survey Albania



Notes: Numerator is the numer of migrants in a given period, denominator the number of migrants in the sample.

A.3 GIS Data for Municipality

All distance indicators are computed in kilometers with the same method: rather than selecting an arbitrary center for each municipality from where to compute distance, we transformed each municipality into rasters with a 30x30 meters grid size.[1] Then, for each cell of each municipality, we computed the straight line distance from the center of the cell to each of the considered geographical points. For each municipality, we then computed the mean distance of all the raster cells it encompasses.

Figure A.2: Albanian municipalities and Italian Television signal

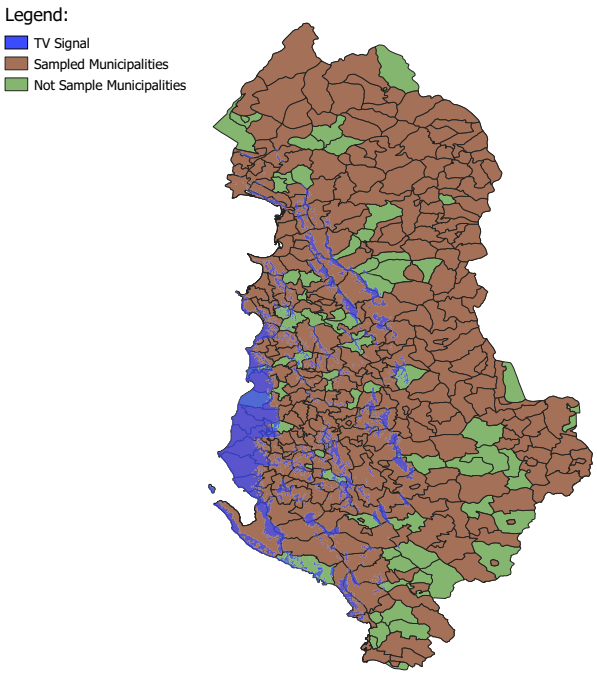


Figure A.2 presents the administrative map of Albania, overlapped with the sampling of the Albania 2005 Living Standards Measurement Survey and signal coverage. The Online Appendix reports the number of observations in the LSMS by district.

A.3.1 LSMS Questions

1. **Foreign language proficiency**

[1] A raster is a geographical object that subdivides a geographical area into cells of equal size.

- Did [NAME] speak English in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
- Did [NAME] speak Italian in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
- Did [NAME] speak Greek in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
- Did [NAME] speak another foreign language in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO

2. **Internal migration**

   - Prior to the current residence, has [NAME] ever lived in a different municipality in Albania? 1 YES, 2 NO
   - Which district and municipality/comuna did [NAME] move from?
   - In what year did [NAME] move to the current residence?
   - Prior to this residence in [MUNICIPALITY/ COMUNA], did [NAME] live in a different municipality/ comuna in Albania? (the loop start over until they track all internal migration history)

3. **Spouse/children away from home**

   - Please list your spouse, if he or she is no longer living in the household, and all the children 15 years old and over who are no longer living in this household. (Include all children of head and/or spouse.)
   - Did [NAME] speak English in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
   - Did [NAME] speak Italian in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
   - Did [NAME] speak Greek in 1990? (1) YES, FLUENTLY , (2) YES, SOME, (3) NO
   - Where does [NAME] currently live? If in Albania, then ask for district and municipality/comuna. If abroad, country and place.
   - In what year did [NAME] move abroad to [COUNTRY]?

4. **Siblings**

- Ask all the questions to the household head, and then to the spouse of the household head. If no spouse, leave the second section blank.
- Please list the first name of up to SEVEN brothers and sisters for both the head of the household and the spouse. Begin with those brothers and sisters living abroad.
- In which country does [NAME] currently live? Indicate the country in which [NAME] spent the most time during the past year
- How many years has [NAME] lived in [COUNTRY]?

5. **TV ownership in 1990**

   - Did your household own any of the following items in January 1990?Colour TV, Black & White TV, Tape player/CD player, Refrigerator, Washing machine, Sewing/knitting machine , Satellite dish, Bicycle.

6. **Education**

   - What is the highest grade you have completed in school? None 0; "8 or 9 years" school 1; Secondary general 2; Vocational 2-3 years 3; vocational 4/5 years 4; University- Albania 5; University- abroad 6; Post-graduate- Albania 7; Post-graduate- abroad 8.

7. **Past migration**

   - Who provided information on where to go and/or how to find work during this most recent migration episode? (MAIN SOURCE) Family/Relatives in Albania; Family/Relatives Abroad; Friends in Albania; Friends Abroad; Previous Personal Experience; Neighbours; TV, Radio, Newspaper or Book; Internet; Other

A.4 Balance Test

Table A.3 reports the results of regressions of age and sex ratio on Italian TV signal using the identification strategy specified in Equation (1.1) and the siblings dataset. We expect to see no effects on age and sex as in our specification being exposed is as good as random. Columns (1) to (4) report results using two measures of signal: share of the municipality exposed to the signal (Signal) and share of urban territory of the municipality exposed in 1986. Concerning the sex ratio we can see there is

no effect of signal on the probability of being a man (Sex ratio). There is, however, an effect on age of two point half year, significant at the 10% level, when using Signal as treatment, while there is no significant effect when using Signal II as treatment. Although an older sample would bias down results (negative correlation between age and migration), we want to rule out the possibility that it is a symptom of issues in our identification strategy. In particular, we show that this result comes from the fact that sets of siblings who have all migrated are absent from the dataset.

Table A.2 shows that in close to a port and to the Greek border areas individuals have significantly higher migration rate and are on average older. As age and migration probability are negatively correlated (-0.2 in our sample) and as the correlation of age between the *listing sibling* and the siblings is extremely high (0.72 in our sample), very low migration migration cost that characterized these areas caused young individuals to be excluded from the sample, via migration of entire set of siblings. Indeed, as we discuss in Data Section 1.5, the siblings dataset contains a sample of Albanians that can be either in Albania or abroad, thus containing migrants, but sets of siblings that all migrated, and only children cannot be a part of the sample. Thus, as close to the port/Greek border areas are differently than average exposed to the signal (Figure 1.1, Table A.2), we observe that Signal affects age although there can not be an effect of our treatment on age. Table A.3 specifications (5)-(6), we show that when we subset for individuals living in 1990 in area farther away from the first quartile of distance to the port (31 km) and first decile of the Greek border (49 km) there is no effect of signal exposure on age.

A.5 Greek Community in Albania

Albania in 1990 was populated by Greek minorities, for many individuals in the survey, Greek is not exactly a foreign language.[2] According to the 1989 Albanian census there were 60 000 Greeks in Albania in 1990, while according to the Greek government they were 300 000. The Communist government recognized 99 villages as *minority zones* in the southern districts of Gjirokastër, Sarandë and Delvina and authorized schooling in both Greek and Albanian for the whole dictatorship period. However, aside from the official recognized minority zones, Greek communities were scattered in many other areas of the country. This is why in Table 2 we control for the Greek community indicators: i) Greek ethnicity ii) Greek

[2]This section is based on the Wikipedia page *Greeks in Albania*: https://en.wikipedia.org/wiki/Greeks_in_Albania

maternal language iii) Use Greek language daily at home iv) Use Greek language with extended family members v) Orthodox religion. In Table A.4, we show that Greek language proficiency in 1990 is correlated with all the Greek community indicators. While language proficiency in Greek is measured in 1990 and the Greek community indicators are available solely for 2005, these indicators are all stables condition that can be assumed to be equal in 1990 and in 2005.

There would be no issue if there were no correlation between signal exposure and Greek settlements in Albania but unfortunately it is not the case. Indeed all *minority zones* are located in districts where signal is available: Delvine (share of signal= .08, share of Greek speaker in 1990=.04), Gjirokastër (share of signal= .01, share of Greek speaker in 1990=.47), Sarande (share of signal= .17, share of Greek speaker in 1990=.22). District F.E. are unable to account for this accidental correlation as the 99 villages could be located precisely in the municipalities exposed to the signal. We consider that finding these 99 villages where both Albanian and Greek was taught would not add much to the research, as Greek community were also scattered across Albania, and it is beyond the scope of the paper.

In Table A.5 we present the regressions (1)-(2)-(3) of Table 1.2 adding as controls Greek community indicators: results are not affected.

A.6 Instrumental Variable Regressions

A.6.1 One-Sample

We can perform an instrumental variable regression on the dataset of children. For this dataset, we have data on both Italian language proficiency in 1990 and on migratory outcomes. We can thus perform a classical IV regression, using Italian television access as an instrument for Italian language proficiency.

The results are presented in Table A.6. When we analyze the entire sample, the first stage of the regression lacks robustness, leading to insignificant results. Upon examining the subset of educated individuals, the first stage proves significant, but the second stage does not. Ultimately, the sample size is insufficient and lacks the precision needed to draw any significant conclusions from this data.

A.6.2 Two-Samples

As outlined in Section 1.7.2, we opted not to conduct two-sample instrumental variable regressions within the primary text body, due to the risk of overestimating the treatment effect. However, we've included this regression analysis in the appendix for the reader's reference. The two stages are represented as follows:

$$IT_{i,m,d} = \alpha_0 + \beta_0 \times Sig_m + \gamma_0 \times Dist_m + \theta_0 \times Geo_m + \sum_{d=1}^{36} \alpha_{0,d} \times Distr_d + \varepsilon_{i,m,d} \tag{A.1}$$

$$MIG_{i,m,d} = \alpha_1 + \beta_1 \times \hat{IT}i,m,d + \gamma_1 \times Dist_m + \theta_1 \times Geo_m + \sum d = 1^{36} \alpha_{1,d} \times Distr_d + u_{i,m,d} \tag{A.2}$$

In equation (A.1), we utilize individuals from the base dataset, where the variable $IT_{i,m,d}$ denotes Italian language proficiency, taking a value of either 0 or 1. In equation (A.2), we draw on individuals from the siblings dataset, with the variable $MIG_{i,m,d}$ representing migration — 1 if an individual migrated and 0 if they did not.

Accurate estimation of the variance-covariance matrix is far from trivial. Inoue and Solon (2010) demonstrates that the variance-covariance matrix needs to be amplified in the case of two-sample instrumental variable regressions. Pacini and Windmeijer (2016) delineates how to perform this adjustment in the presence of heteroskedastic errors. To date, no formal proof has been derived for clustered standard errors. We have applied the methodology as described by Etgeton (2018), inspired by Pacini and Windmeijer (2016), but not yet analytically proven.

The results are presented in the corresponding table. Due to the notably high estimated standard errors in relation to the small sample size, the results are inconclusive. The standard errors in the second stage are simply too excessive to yield any significant conclusions.

A.6.3 Additional Results

In this section we present additional results absent from the main body of the paper. In particular, we show the null effect of signal exposure on presumably low skill individuals and the heterogeneity of the effect over the family dimension and the housing dimension. (1)-(2) show the absence of an effect on migration of low skilled individuals. (3) to (5) confirm that the effect on migration is decreasing

in family size.[3] (6)-(7) show that there are no effects of signal exposure for individuals who were living in smaller housing in 1990 (2nd and 3rd quartile for specification (6) and 4th quartile for specification (7)). We confirm that there is no effect for *low skilled* individuals as proxy by sibling education, family and housing dimension.

[3]For sake of brevity we do not show family dimension higher than 5. The coefficient steadily decline with family dimension.

Table A.1: Italian TV Shows Preferences of Albanians Migrants

	Italian TV Shows Preferences of Albanians Migrants		
	Obs. (1)	Share (2)	Type (3)
Domenica In	183	25%	Entertainment
Fantastico	92	13%	Entertainment
Piacere Raiuno	86	12%	Entertainment
Domenica sportiva	84	11%	Entertainment
Creme Caramel	35	5%	Entertainment
Quark	34	5%	Entertainment
Sanremo	30	4%	Entertainment
La Piovra	25	3%	Entertainment
Lunedi film	23	3%	Entertainment
Tg1	21	3%	Information
Mercoledì sport	19	3%	Entertainment
Big	17	2%	Entertainment
Tg1 7	13	2%	Information
Discoring	13	2%	Entertainment
Speciale Tg1	12	2%	Information
Linea Verde	12	2%	Entertainment
Viaggio intorno all uomo	10	1%	Entertainment
Colpo Grosso	10	1%	Entertainment
Telemike	9	1%	Entertainment
Notte Rock	7	1%	Entertainment

Source: Data derived from Dorfles and Gatteschi (1991).

Table A.2: Distance to Ports and Greek Border: Signal, Migration, and Age

	Distance to Ports		Distance to Greece	
Variable	$\leq$ 31km	$>$ 31km	$\leq$ 49km	$>$ 49km
Migration	0.210	0.150	0.210	0.160
Age	49.4	46.9	49.2	47.6
Signal	0.180	0.060	0.020	0.110

Notes: We display the average share of individuals being abroad, the average age, the average signal exposure for individuals in/outside the first quartile of distance to the nearest port (30.461 km) and in/outside the first decile of distance to the Greek border (48.834 Km).

Table A.3: Balance test: Age and Sex Ratio

	Full Sample				Restricted Sample	
	(1)	(2)	(3)	(4)	(5)	(6)
	Age	Sex ratio	Age	Sex ratio	Age	Sex ratio
Signal	2.434*	-0.0253			0.441	-0.0573
	(1.254)	(0.0232)			(2.531)	(0.0450)
Signal II			1.497	-0.0208		
			(1.175)	(0.0197)		
Observations	27666	27666	27666	27666	18985	18985
Clusters	310	310	310	310	180	180
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness						

Notes: The table reports OLS estimates of the effect of exposure to Italian Television on the probability of being a man (Sex ratio) and an individual's age. All specifications exploit the *siblings* dataset. Specifications (1) to (4) exploit the full sample, while specifications (5) and (6) restrict the sample to individuals that lived in 1990 farther away from 30.461 km from the closest port (first quartile of distance to the port) and farther away from 48.834 km from the Greek border (first decile of distance to Greece). Signal is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Signal II is the share of the urban area in the municipality exposed in 1986. Clustered standard errors in parentheses. Standard errors are clustered at the Municipality level.

Table A.4: Correlation between Greek Proficiency in 1990 and Greek Community Indicator

	Proficient in Greek in 1990
Variable	Correlation coefficient
Greek Ethnic Group	0.51
Greek Maternal Language	0.52
Greek spoken daily at home	0.47
Greek spoken with extended family	0.36
Orthodox religion	0.22

Source: 2005 Living Standard Measurement Survey, World Bank and INSTAT. Base dataset.

Table A.5: Italian television effect on foreign language proficiency: controlling for Greek community indicators

	Italian	English	Other
	(1)	(2)	(3)
Signal	0.0635*	0.0118	0.00191
	(0.0334)	(0.0181)	(0.0231)
Obs.	11040	11040	11040
Clusters	322	322	322
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators			

Notes: The table reports OLS estimates of the effect of exposure to Italian television on foreign language proficiency in 1990. (1)-(4) exploit the sample of the LSMS surveyed individuals. The dependent variable is the reported capability of speaking Italian, English, Other (category any other language). The main explanatory variable, Signal, is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Clustered standard errors in parentheses. Standard errors are clustered at the municipality level.

Table A.6: IV Regression Results

	(1)	(2)
	First Stage	
Signal	0.110	0.481**
	(0.071)	(0.213)
	Second Stage	
Italian Proficiency	0.053	0.717
	(.620)	(0.464)
Sample	All	H. Skill
Obs.	4,714	425
Clusters	256	104
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators		

Notes: The dependent variable is migration, the endogenous variable Italian proficiency, and the instrument signal exposure (see section 1.5 for details). Both stages are linear probability model (see section 1.6). Standard errors are clustered at the municipality of living in 1990 level.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.7: Two samples 2SLS

	Dependent variable:		
	Italian Proficiency	Abroad	
	(1)	(2)	(3)
Signal	0.070**		
	(0.033)		
Italian Proficiency		−0.063	2.455
		(1.282)	(29.188)
Sample	All	All	H. Skills
Obs.	11,040	27,666	2,153
Clusters	322	310	128

Notes: We use two-sample 2SLS IV regression. The dependent variable, migration, is only in the siblings dataset, the endogenous variable Italian proficiency only in the base dataset, and the instrument signal exposure in both (see section 1.5 for details). Both stages are linear probability model (see section 1.6). Standard errors are clustered at the municipality of living in 1990 level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: The effect of Italian Television exposure on migration decision: Other Results

	Low Skilled		Family Dimension: # Children			Housing Dimension	
	Abroad	Italy	Abroad	Abroad	Abroad	Abroad	Abroad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Signal	0.0103	0.0157	0.248**	0.0502	0.0435	-0.0140	0.0343
	(0.0307)	(0.0172)	(0.111)	(0.0459)	(0.0354)	(0.0414)	(0.0787)
Obs.	25513	25513	517	5926	10986	5922	2442
Clusters	302	302	172	268	287	259	143
Sample	L. Skill	L. Skill	≤ 2	≤ 5	≤ 6	Quart. 2-3	Quart. 1
Controls: District F.E., Distance to Italy, Distance to transmitter, Distance to port, Elevation, Ruggedness, Distance to Greece, Greek community indicators							

Notes: The table reports OLS estimates of the effect of exposure to Italian television on: (1) probability to be abroad and (2) probability to be in Italy on low skilled individuals (*listing brother* education less equal than secondary education), (3)-(4)-(5) probability to be abroad given the number of children of the family an individual was raised in, (6)-(7) probability to be abroad given housing dimension. All specifications use subset of the *siblings* dataset. The main explanatory variable, Signal, is the share of the municipality area (where an individual i was living in 1990) exposed to Italian television signal in 1990. Clustered standard errors at the municipality level in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix B

Second Appendix

B1 Descriptive Statistics

Descriptive statistics. Table B.1 presents the descriptive statistics of our sample. The share of individuals who graduated from university (with a master degree), from a business or an engineering school has increased substantially. We also observe that their average real log wage has decreased. In contrast, the share of individuals with less than a high-school degree has decreased, but their average log-wage has increased. In general, we observe that the average real monthly wage of full-time employees increases with education. More interestingly, we remark that there is an increase in the range of the share of individuals working full-time across education degrees. Whereas in 1998, the share of individuals working full time varied from 62% (low education level) to 68% (high education level), in 2010, the same share varied from 46% among the individuals without a high-school degree to 76% for individuals who graduated from a business (or engineering) school.

B2 Likelihood

We derive the model's likelihood function. Individuals are indexed by $i = 1, \ldots, N$. Recall that the log-wage w_{it} is observed in a subset T_i of periods. The probability density of w_{it}, conditional on observed characteristics and latent type k, is denoted as follows,

$$p(w_{it}|x_{it}, X_i, h_i, k) = f_k(\varepsilon_{itk}), \tag{B.1}$$

where f_k is the pdf of a normal distribution with mean 0 and standard deviation σ_{wk} and ε_{itk} is defined by equation (2.2). Now, denote $w_i = (w_{it})_{t \in T_i}$ and $x_i = (x_{it})_{t_i \in T_i}$. We have

$$\Pr(w_i|x_i, X_i, h_i, k) = \prod_{t \in T_i} p(w_{it}|x_{it}, X_i, h_i, k).$$

The probability of observing an employment rate conditional on past observed employment rates, exogenous characteristics, education level h and latent type k can

Table B.1: DESCRIPTIVE STATISTICS:
AVERAGE EMPLOYMENT RATE AND FULL-TIME REAL LOG-WAGES

Cohort	Education	Individuals	Obs.	Average Employment	Average FT Real Log-Wage
1998	Less than High School	3026	35285	0.65	7.16
	High-School Degree	1815	19216	0.64	7.23
	Some College, Bachelors	2038	20902	0.64	7.39
	Masters	203	1826	0.64	7.72
	Bus. and Engin. Sch. Deg.	301	2777	0.69	7.85
	All	7383	80006	0.65	7.28
2004	Less than High School	1601	20475	0.61	7.23
	High-School Degree	1406	16038	0.65	7.27
	Some College, Bachelors	1628	16580	0.68	7.39
	Masters	470	4363	0.70	7.65
	Bus. and Engin. Sch. Deg.	395	3461	0.73	7.80
	All	5500	60917	0.65	7.36
2010	Less than High School	870	10855	0.50	7.22
	High-School Degree	917	10538	0.58	7.27
	Some College, Bachelors	964	9554	0.65	7.40
	Masters	419	3869	0.67	7.62
	Bus. and Engin. Sch. Deg.	351	2673	0.76	7.78
	All	3521	37489	0.60	7.39

be written as follows:

$$P_k(e_{it}|X_i,x_{it},h_i) = \Pr(e_{it}|X_i,x_{it},h_i,k) = \prod_{g=1}^{G}\left[F(\mathbf{c}_{g+1,k}-\rho_{itk}) - F(\mathbf{c}_{g,k}-\rho_{itk})\right]^{Q_{itg}}, \tag{B.2}$$

where

$$Q_{itg} = \begin{cases} 1 & \text{if } e_{it} = \mathbf{e}_g \\ 0 & \text{otherwise} \end{cases},$$

and F is the cumulative distribution function of the standard normal distribution. Finally, we denote the probability of choosing education level h_i conditional on observable characteristics Z_i and latent type k as follows,

$$\Lambda_k(h|Z_i) = \frac{\exp(\mathbf{v}_{ihk})}{\sum_{j=1}^{H}\exp(\mathbf{v}_{ijk})}. \tag{B.3}$$

Let now y_i denote the vector of outcomes of individual i, namely, observed wages w_{it}, observed employment rates e_{it} and the observed education (*i.e.*, highest degree) h_i. Let E_i be the set of periods during which i's employment rate e_{it} is observed. Let $e_i = (e_{it})_{t\in E_i}$. Recalling that $x_{it} = \sum_{\tau=1}^{t-1} e_{i\tau}$, we can write the conditional probability

of e_i as follows,

$$\Pr(e_i \,|\, X_i, h_i, k) = \prod_{t \in E_i} \Pr(e_{it} | x_{it}, X_i, h_i, k), \tag{B.4}$$

where $x_{i\tau} = 0$ if i enters the labor market at time τ for the first time.

Then, we can write the contribution to likelihood of an individual i with type k as,

$$\begin{aligned} L_{ik} = L_{ik}(y_i|X_i) &= \prod_{t \in T_i} p(w_{it}|x_{it}, X_i, h_i, k) \prod_{t \in E_i} \Pr(e_{it}|x_{it}, X_i, h_i, k) \Pr(h_i|X_i, k) \\ &= \left(\prod_{t \in T_i} f_k(\varepsilon_{itk}) \right) \left(\prod_{t \in E_i} P_k(e_{it}|x_{it}, X_i, h_i) \right) \Lambda_k(h_i|Z_i), \end{aligned} \tag{B.5}$$

where ε_{itk} is defined by equation (2.2).

Now, integrating over latent types k, the contribution to likelihood of individual i can be written,

$$L_i(y_i|X_i) = \sum_{k=1}^{K} p_k L_{ik}(y_i|X_i), \tag{B.6}$$

The model Likelihood is $L = \prod_{i=1}^{N} L_i$, so that the Log-Likelihood is

$$\ln L = \sum_{i=1}^{N} \ln \left[\sum_{k=1}^{K} p_k L_{ik} \right], \tag{B.7}$$

The *posterior* probability that individual i is of type k is denoted p_{ik}; it can be expressed with the help of Bayes' rule and the likelihood, as follows,

$$p_{ik} = \Pr(k|X_i, y_i) = \frac{p_k L_{ik}}{\sum_{j=1}^{K} p_j L_{ij}}. \tag{B.8}$$

The posterior probabilities are a crucial ingredient in many useful computations. For all $k = 1, \ldots, K$, we have,

$$p_k = \frac{1}{N} \sum_{i=1}^{N} p_{ik}. \tag{B.9}$$

It is easy to see that the latter equation is a necessary condition for likelihood maximization[1], and it follows that this relation between priors p_k and posteriors p_{ik} holds when we use their numerical, estimated values.

B3 Simulations

The simulations can be decomposed in a few steps.

Step 1.

[1] Indeed, it is equivalent to $\partial \ln L / \partial p_k = 0$, for $k = 2, \ldots, K$ where we set $p_1 = 1 - \sum_{k=2}^{K} p_k$.

1(a). We first recursively simulate the employment level $\tilde{e}_{itk}$ for each (i,t), and each k, $t = 1,\ldots,T$, $k = 1,\ldots,K$ and $T = 84$.

We start with $t = 1$ and then increment. We initialize experience by setting $\tilde{x}_{i1k} = 0$. We draw a random number $\tilde{\zeta}_{itk}$ for each (itk), with $\tilde{\zeta}_{itk} \sim \mathcal{N}(0,1)$. Then, we use the ordered probit as estimated by ML. More precisely, if it happens that

$$c_{gk} - \rho_{itk} \leq \tilde{\zeta}_{itk} \leq c_{g+1,k} - \rho_{itk},$$

where ρ_{itk} is given by 2.5 above, then we set $\tilde{e}_{itk} = e_g$. To compute ρ_{itk} we use $x_{it} = \tilde{x}_{itk}$ for $t > 1$. Recall that $e_g \in \{0,.3,.5,.8,1\}$.

1(b) Compute the accumulated experience $\tilde{x}_{itk} = \sum_{\tau<t} \tilde{e}_{itk}$, with $\tilde{x}_{i1k} = 0$.

Step 2. Given the sequences $(\tilde{e}_{itk}, \tilde{x}_{itk})$, we compute a sequence of expected log-wages for each (i,t,k) (no need to draw a random shock here). Using the estimated values of the parameters, we set, for each (i,t,k),

$$\tilde{w}_{itk} = \mathbb{E}[w_{itk}|\tilde{x}_{itk}, X_i] = \alpha_{0k} + \beta_{0k}\tilde{x}_{it} + \sum_{h=1}^{H} \gamma_{0hk}\chi_h(i) + X_i\eta_{0k}.$$

Step 3. Given the simulated sequences $(\tilde{e}_{itk}, \tilde{w}_{itk}, \tilde{x}_{itk})$ we can now compute the discounted expected earnings during the periods $t \in \{1,\ldots,T\}$. We choose a discount factor δ and for every (i,k), we compute

$$\tilde{W}_{ik} = \frac{(1-\delta)}{(1-\delta^T)} \sum_{t=1}^{T} \delta^{t-1} \tilde{e}_{itk} \exp(\tilde{w}_{itk}).$$

$\tilde{W}_{ik}$ has the dimension of monthly earnings[2]

Then, we compute the weighted arithmetic mean, using the estimated probabilities p_{ik}. For each type k, we compute,

$$H_k = \frac{\sum_{i=1}^{N} \tilde{W}_{ik}\hat{p}_{ik}}{\sum_{i=1}^{N} \hat{p}_{ik}}.$$

We can also compute expected-discounted values conditional on a degree h. So, we define $I(h) = \{i|h_i = h\}$ and we compute,

$$H_k(h) = \frac{\sum_{i\in I(h)} \tilde{W}_{ik}\hat{p}_{ik}}{\sum_{i\in I(h)} \hat{p}_{ik}},$$

which measures the average expected-discounted earnings of a type k, knowing the degree h. This type of conditioning can be performed with any other subsample (for instance, the sons of executives, or the sons of executives with degree h, etc.).

[2]We choose a yearly discount rate of 0.9845. This corresponds to a monthly discount rate $\delta = 0.9987$. $\tilde{W}_{ik}$ is a weighted average of expected monthly earnings with weights $(\delta^{t-1}(1-\delta))/(1-\delta^T)$.

[3]

B4 Full tables: Wage Equation

Table B.2: Wage equation. Returns to experience

Type		1	2	3
Experience ×				
1998 cohort	Below High-school degree	0.0022	0.0028	0.0047
		(.00008)	(.00008)	(.0002)
	High school degree	0.0030	0.0034	0.0059
		(.00009)	(.0001)	(.0003)
	Some College and Bachelors	0.0041	0.0036	0.0068
		(.00009)	(.0001)	(.0003)
	Masters	0.0039	0.0043	0.0062
		(.0003)	(.0003)	(0006)
	Bus. Engin. School degree	0.0026	0.0038	0.0098
		(.0003)	(.0003)	(.0005)
2004 cohort	Below High-school degree	0.0016	0.0021	0.0031
		(.0001)	(.0001)	(.0003)
	High-school degree	0.0019	0.0024	0.0046
		(.0001)	(.0001)	(.0003)
	Some College and Bachelors	0.0026	0.0028	.0054
		(.0001)	(.0001)	(.0003)
	Masters	0.0044	0.0040	0.0059
		(.0002)	(.0002)	(.0004)
	Bus. Engin. School degree	0.0040	0.0035	0.0063
		(.0002)	(.0002)	(.0005)
2010 cohort	Below High-school degree	0.0024	0.0021	0.0040
		(.00019)	(.0002)	(.0005)
	High-school degree	0.0028	0.0027	0.0047
		(.00013)	(.0002)	(.0004)
	Some College and Bachelors	0.0031	0.0035	0.0043
		(.00013)	(.0001)	(.0004)
	Masters	0.0041	0.0033	0.0056
		(.00026)	(.0002)	(.0004)
	Bus. Engin. School degree	0.0029	0.0050	0.0053
		(.00023)	(.0003)	(.0005)

[3]These computations can be improved, if needed, by simulating employment and wage trajectories several times in the same fashion and then taking the simple arithmetic averages of all simulated values of H.

Table B.3: Wage equation. Returns to education

Type		1	2	3
1998 cohort	High-school degree	-0.006	0.04	.12
		(.006)	(.007)	(.015)
	Some College and Bachelors	0.072	0.20	0.27
		(.006)	(.008)	(.014)
	Masters	0.59	0.70	0.17
		(.015)	(.019)	(.028)
	Bus. Engin. School degree	0.55	0.64	0.36
		(.018)	(.012)	(.027)
2004 cohort	High-school degree	0.014	0.016	0.03
		(.006)	(.01)	(.020)
	Some College and Bachelors	0.038	0.11	0.14
		(.007)	(.009)	(.019)
	Masters	0.14	0.30	0.35
		(.011)	(.014)	(.026)
	Bus. Engin. School degree	0.25	0.43	0.46
		(.013)	(.011)	(.029)
2010 cohort	High-school degree	0.02	0.04	0.007
		(.009)	(.014)	(.027)
	Some College and Bachelors	0.06	0.12	0.256
		(.009)	(.011)	(.028)
	Masters	0.14	0.29	0.39
		(.013)	(.016)	(.030)
	Bus. Engin. School degree	0.56	0.58	0.20
		(.014)	(.017)	(.030)

B5 Multinomial Logit; Full Estimation Results

Table B.4: Wage equation. Controls

Type	1	2	3
2004 cohort	0.09	0.09	0.15
	(.006)	(.007)	(.016)
2010 cohort	0.07	0.10	0.15
	(.008)	(.01)	(.023)
Father is a professional	0.012	0.019	0.06
	(.003)	(0.004)	(.006)
Peri-urban	-0.006	-0.010	-0.039
	(.003)	(.003)	(.007)
Rural	-0.018	-0.022	-0.025
	(.002)	(.003)	(.006)
Unemployment rate	0.002	-0.015	-0.020
	(.001)	(.001)	(.003)
Constant	6.98	7.26	7.31
	(.01)	(.01)	(.028)

B6 Online Appendix: Ordered Probit; Full Estimation Results

B7 Online Appendix: Results obtained with the Elastic Net Method

Table B.7 reports the results of the elastic net method applied to the most-likely-type indicators. The coefficients of the explanatory variables selected by the algorithm appear in the table (otherwise, the entry is blank); these variables are significant. There are four groups of three columns, one for each type in each group of three. The first three column groups correspond to the three cohorts, 1998, 2004, 2010. The last group of three columns reports results obtained when the three cohorts are stacked. There are indicators of the father's and the mother's occupation and indicators of the region of origin listed in the bottom half of the table. Grade repetition; rural origin; parents are farmers; mother is a graduate; Corsica; South-West of France (Occitanie and Aquitaine); West Indies and Islands; Paris are the most salient indicator variables: this is not particularly surprising.

B8 Online Appendix: A Preliminary Analysis Using Standard Econometric Methods

B8.1 The devaluation of degrees.

Estimation in sub-samples of students holding the same degree

We now present some preliminary results obtained with the unbalanced panel stacking the surveys of 1998, 2004 and 2010. The details are given in several appendices. We first consider the most classical log-wage regression. Log-wages are regressed on potential experience and dummies interacting the education level with the cohort. Potential experience is denoted z_{it}. Potential experience is defined as the number of months elapsed since the individual left the educational system. In essence, we studied variants of the following regression:

$$w_{it} = a + \sum_k \mathbb{1}_k(i)(b_k + c_k z_{it} + d_k z_{it}^2) + \varepsilon_{it}, \tag{B.10}$$

where (a, b_k, c_k, d_k) are parameters, ε_{it} is a random error with a zero mean and $b_1 = 0$. Index k can denote the education level, the cohort (the survey), gender, or any interaction of the three. Variable $\mathbb{1}_k(i)$ is a dummy equal to 1 when i has characteristic k.

We start with a test showing the devaluation, on average, of higher education degrees (we set $c_k = d_k = 0$). Table B.8 gives the results of 5 simple sub-sample regressions, one for each education level, of log-wages on dummies indicating the cohort, estimated by OLS on pooled data, without any control for experience. Taking the 1998 cohort as a reference, we find a significant devaluation for some degrees, the drop in average real wages being particularly clear, of the order of -9% for the Master's (M2) and -6% for the Engineering and Business school degrees, between the 1998 and 2010 cohorts. In contrast, the corresponding results for attainment levels below or equal to high-school graduation (*i.e.*, below the French *baccalauréat*) did not suffer any devaluation. On the contrary, in these categories, we see only real-wage increases. This striking difference can be attributed to minimum-wage regulations. Indeed, the real-value of the minimum wage rose by 26% between 1992 and 2012. This substantial growth protected the less skilled working full-time from the devaluation observed at the other end of the hierarchy of degrees. Of course, these results do not take care of the fact that years of education are an endogenous variable; our coefficients are not average treatment effects, only ATTs.

The Decline of Returns to Experience

In fact, we can show that a substantial part of the observed devaluation takes the form of a decrease in the returns to potential (or effective) experience during the first 7 years of career. If we add a control for potential experience in linear form

(*i.e.*, now allowing for $c_k > 0$ and keeping $d_k = 0$), we find that the returns to potential experience decreased with time (see Table B.9 in Appendix B9). Returns to potential experience are of the order of 3% per year. They increase with the number of years of education. Appendix B9 gives the within-group estimates of returns to potential and effective experience obtained with our data.[4]

B8.2 Unemployment, the Business Cycle and the Supply of Graduates

A look at the overall unemployment rate is needed, for the observed devaluation of degrees could entirely be due to a business-cycle effect. We consider the national unemployment rate, which is correlated with GDP. Variations of the unemployment rate do have an impact on real wages in our dataset. The real wages are sticky and mildly procyclical. In any case, we will control for the variation of national unemployment to capture a business-cycle effect. A discussion of this point is developed in Appendix B10.

B9 Online Appendix: The Returns to Experience. Fixed-effects, *Within* Estimators

To summarize the information conveyed by many regression coefficients, we can plot the wage $W = e^w$ as a function of potential experience and potential experience squared for top and bottom degrees. Coefficients of the Mincer regression (B.10) are all precisely estimated. We used these coefficients to draw the curves of Figures B.1 and B.2. We see the devaluation of top degrees in the case of male students on Fig. B.1: the dotted line representing the Mincer curve of the 2010 cohort starts slightly above the other two, but after a year and a half of career, the young workers of the 1998 cohort were better off. This result is no longer true for the lowest degrees on Fig. B.2.[5] The latter result can be almost entirely attributed to the growth of the French minimum wage. These pictures are of course not statistical tests; they should be viewed as a convenient representation of estimated coefficients, but with our data, the returns to experience are very precisely estimated (see below).

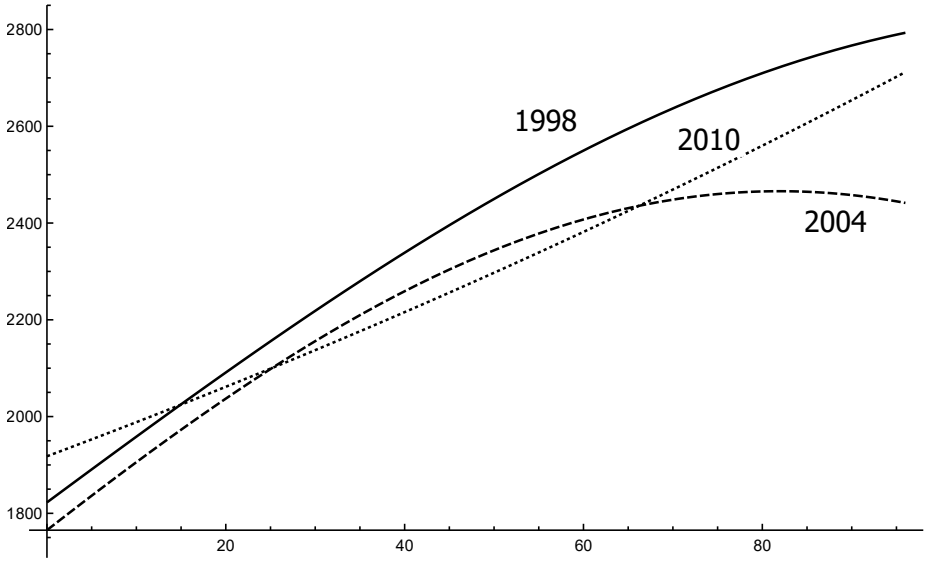
Using the panel structure of our data, we can easily obtain fixed-effects, within-group estimates of returns to experience, either potential (*i.e.*, z_{it}) or effective (*i.e.*, x_{it}). We assume that the endogeneity of experience stems from additive individual effects (*i.e.*, the so-called fixed effects).[6]

[4] A comparison of OLS and *within* estimators shows that OLS estimates of returns to effective experience by means of regression (B.10) are upward biased. If the bias is small for basic secondary degrees, it is in contrast substantial in the case of higher-education degrees.

[5] Appendix **??** shows that the picture is somewhat different for young women, whose salaries resisted devaluation much better: after 5 years (*i.e.*, 60 months), the 1998 curve catches up the 2010 curve. In Appendix **??**, Fig. **??** shows that the less educated women did not experience any devaluation during our 20 year period.

[6] Error terms can be written $\varepsilon_{it} = u_i + v_{it}$ where v_{it} have a zero mean and are independent of explanatory variables and u_i. Terms u_i are individual effects depending on i that do not vary with

Figure B.1: **Men. Masters and 'Schools' Degrees**



Comparison of returns to education and experience of male workers holding *Master's, business and engineering school degrees* in three 7-year Generation surveys, 1998, 2004, 2010. Months of potential experience are on the x-axis; monthly real wages (2013 euros) are on the y-axis.

We again study the wages of full-time jobs, but we use all employment spells to compute an individual's effective experience (as explained above). Effective experience is potentially highly endogenous, because individuals with the best characteristics on the labor market also accumulate more experience. OLS estimates of the returns to effective experience should therefore overestimate these returns. This is what we find.

Table B.9 permit a comparison of estimated returns of both potential and effective experience in the three cohorts and for three aggregate levels of educational attainment. The estimates of c_k are monthly returns to experience.[7] We obtain yearly returns, denoted γ_k, with the formula, $\gamma_k = (1+c_k)^{12} - 1$. All c_k coefficients are significant at the .1 percent (*i.e.*, 10^{-3}) level.[8] Table B.9 gives coefficients γ_k, where c_k is estimated with two different methods: by OLS on pooled data and by the fixed-effects, *within* estimator (FE).

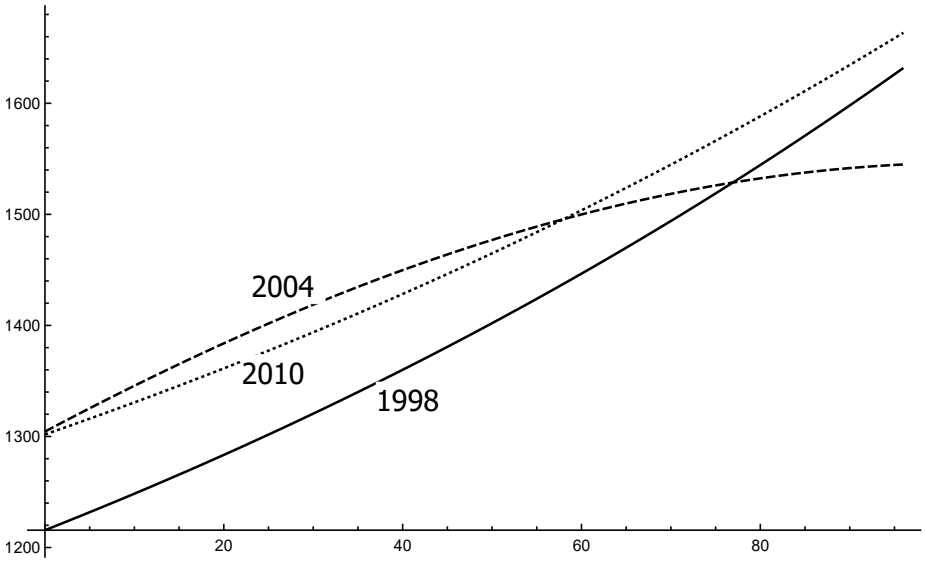
The first striking fact is that returns to experience are substantial, with values ranging from 2% to 6% per year. The OLS returns to potential experience seem to be only slightly biased (if we compare the estimated coefficients with the corresponding fixed-effects coefficients). In contrast, as expected, the OLS returns to effective experience are biased upwards, and all the more since the attainment level is high. Returns to experience typically increase with the education level, in all co-

time t. The within estimator will then produce unbiased estimates of c_k and d_k.

[7] In Appendix **??**, Table **??** gives the equivalent results for the subsample of women.

[8] We assumed $d_k = 0$ here.

Figure B.2: **Men. High-School Degree and Less**



Comparison of returns to education and experience of male workers with an educational achievement lower than or equal to high-school graduation *i.e.*, the French *baccalauréat* in three 7-year Generation surveys, 1998, 2004, 2010. Months of potential experience are on the x-axis; monthly real wages (2013 euros) are on the y-axis.

horts. But the most important feature of Table B.9 is that returns to experience fell between 1998 to 2004, and they fell more for the highest degrees of attainment.

B10 Online Appendix: Impact of the Business Cycle. Variations of the National Unemployment Rate

In France, from 1998 to 2002, the national unemployment rate dropped from 10.3% to 7.9%. Then, unemployment grew again and reached 9% in 2010 and 10.4% in 2015, as shown by Figure B.3. In spite of the relatively better macroeconomic conditions of 1998-2001, the devaluation of higher-education degrees is already visible when we compare the surveys of 2004 and 1998, as shown by Table B.8. The first years of *Génération* 2004 are characterized by a relatively small national rate of unemployment. In the middle of the 7-year period, *i.e.*, after 2007, the effects of the great recession started to be felt, slowing down wage increases. The 2010 survey is characterized by a relatively higher unemployment rate reaching 10%. In the quarters following 2010, France came back to the high unemployment situation of the beginning of 1998. In spite of the swings of macroeconomic unemployment, during the entire period, we observe the downward trend of the real wages of university and engineering-school graduates, while at the same time, the real wages of young workers with less than a high-school degree grew. There is no simple explanation of the evolution of real wages in terms of business-cycle fluctuations.

Figure B.3: **Unemployment rate; France, 1992-2019**

The jobs of educated workers are stable as compared to that of unskilled labor. It could still be true that the variation in real wages is caused by the business cycle because real wages are more flexible at these levels of education.[9] Yet, as we will see, other factors are likely to be responsible for (most of) the observed devaluation. We checked that variations of the unemployment rate do have a significant impact on real wages. Real wages are sticky and hence mildly procyclical. But the estimated returns to education by degree and cohort (or zero-experience wages), taking the 1998 high-school dropouts as a reference, are robust to the introduction of a control for variations of national unemployment.[10] These estimates are particularly stable for higher levels of educational attainment. Indeed, if we control for the variation of unemployment, we still find a drop of around 5% in the return of a 2-year Master's degree, relative to 1998.[11] Returns to experience also seem to be procyclical. We conclude that we should control for variations of the national unemployment rate, but that the devaluation phenomenon is not closely related to the business cycle.

[9] In France, the minimum wage legislation is the main cause of rigidity at lower levels.

[10] In Online Appendix B11, we show the result of a log-wage regression in which we control for the variation of the macroeconomic unemployment rate; see Table B.10.

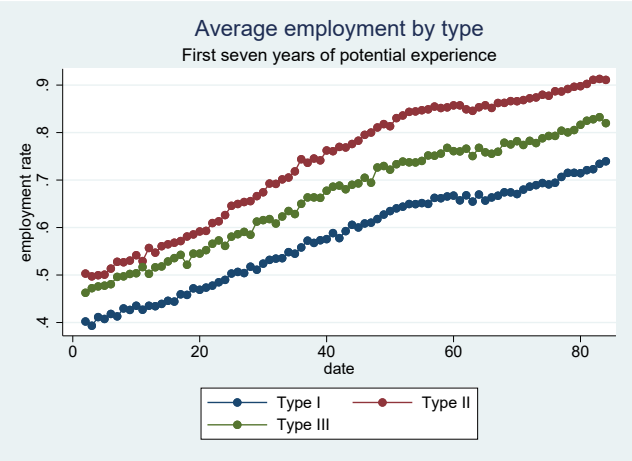
[11] See Online Appendix B11, Table B.10.

B11 Online Appendix: Adding a Control for the Business Cycle

Table B.10 displays a subset of estimated coefficients from a Mincer-type equation where log-wages are regressed on degree indicators interacting with gender and the cohort, plus terms where potential experience interacts with gender, degree and the cohort, and finally, on the rate of growth of the overall, macroeconomic unemployment rate (more precisely, the variation of the logarithm of the national rate of unemployment). We compare the results obtained with or without a control for the variation of the rate of unemployment.[12] A glance at Table B.10 shows that the variation of unemployment, measuring the effect of the business cycle, has a weak impact on men's skill premia at zero experience. Yet, the unemployment variable has a significant impact on wages. These results how the robustness of the devaluation of university degrees emphasized above.

[12] This measure of unemployment variation is used here because it yields better results than the French GDP or the rate of variation of the French GDP.

Figure B.4: Employment Rates by Type: Simulations



B12 Online Appendix. Differences in Employment Rates by Type

Our estimates of the employment equation exhibit a few other interesting properties. The probability of full-employment generally decreases with the cohort, and particularly for types 1 and 2. In a certain sense, this contributes to the devaluation of degrees. The probability of a higher rate of employment typically increases with educational achievement. The impact of effective experience on the probability of employment is always positive and significant, showing the existence of a virtuous circle of employment (employment today begets more employment in the future) and this effect is higher for the 2010 cohort than for older cohorts; it seems also stronger for type 2 than for types 1 and 3.

Can we provide an intuition for the reason why Type 2's jobs are so stable? A first reason has to do with a possible contrast between the public and private sectors (because public jobs are typically much more stable than private sector equivalents in France). Indeed, Table B.11 shows that Type 2 is slightly more frequent in the public sector (85% of the Type 3 are in the private sector, as compared to only 78% for Type 2). So, this seems to be an important difference between Type 2 and Type 3.

Firm size could be an even more important source of job stability. The Generation Surveys give indications on the size of the employing firm during each employment spell. Table B.12 shows that Type 1 is evenly distributed in the three size categories, while between 50 and 60% of Types 2 and 3 are in large firms. Table B.12 does not show a real difference between Type 2 and Type 3 when it comes to firm size. We conclude that observable characteristics of jobs and employing firms help understanding the specific visible features of Type 2, but only to a certain extent. We'll come back below to the fact that types are not well explained by

omitted controls.

We can check that the average number of employment spells of Type 2 is smaller than two while Types 1 and 3 have an average number of spells greater than three. This is shown on Table B.13 and confirms the greater job stability of Type 2.

Finally, a look at simulated employment and experience paths clearly shows the differences between types. Simulations confirm that Type 2 has the highest employment rate, as shown by Figure B.4.

B13 Impact of Family Background by Type

When the father is a professional, the probability of reaching a higher education level is significantly increased for all types — the probability of reaching the top levels is markedly increased.

On the role of an educated father (*i.e.*, "father is a professional") we can provide more indications. Table B.14 gives the estimated percentage of individuals whose father is a professional (knowing that this category includes mainly educated fathers: teachers, engineers, doctors, executives etc.). Table B.14 shows that the proportion of professional fathers is increasing with time (this is due to the fact that the years of education of the population are increasing). In addition, we see that educated fathers are more prevalent among Type-3 individuals. But the professional father is far from perfectly correlated or predicted by the type.

Table B.15 is particularly striking. It gives the estimated probabilities of reaching (choosing), the various educational levels, conditional on cohort, type and the fact that the 'father is a professional'. This Table is all more striking if we compare it to the unconditional equivalent, that is, Table 2.7. As time passes, the sons of professionals have been deserting the lowest educational levels and invading the highest levels, in particular, the Type 3s and, to a lesser extent, the Type 2s with a professional father, the most impressive "invasion" being that of Master programs by individuals with this family background.

B14 Online Appendix: Choice of the Number of Types K; Robustness

Until now, we estimated the model with three types, but the number of types is a choice that must be justified. The choice of a number of types K is in principle not easy. We devote the next subsection to this question. It happens that 3 types seems to be the right choice. The model has been estimated with panel data stacking three surveys. But would the results be different had we estimated the same model three times separately with the help of each survey? It happens that the results do not change much if we try this form of sub-sample estimation.

B14.1 Number of Types

The question of the number of types is crucial because the set of types provides a model of the unobservable factors generating the well-known endogeneity problems: mainly the endogeneity of education and experience in the wage and employment equations.

The difficulty comes from the well-known fact that the log-likelihood of the model with K types, denoted $\mathscr{L}(K)$, is typically increasing and concave: an additional type will always lead to some improvement of $\mathscr{L}(K)$, but with decreasing marginal values. If K is too small, the types are themselves heterogenous melting

pots of individuals. If K is too large, there is a risk that the types do not represent real individuals but are just improving the approximation of the distribution of wages, education and employment by a finite mixture of normal distributions. We know that, in essence, any distribution can be approximated by a mixture of normals, to any desired degree of precision, and in our case, a large K may simply be a form of over-fitting.

To choose the number of types K, we in fact combine several criteria. The usual criteria penalizing the likelihood for a high number of parameters, the Akaike and the Bayesian Information Criteria (resp. AIC and BIC, see Akaike (1974), Schwarz (1978)) will in principle reach a minimum for some value of K, but are not well adapted to the choice between K and $K+1$.[13] AIC tends to overestimate the correct number of components (AIC pushes towards over-fitting). BIC corrects for these difficulties but tends to underestimate K.[14] These criteria are useful, but they do not measure the quality of classification. So we use other criteria, based on *entropy* and penalizing the fact that types are difficult to distinguish.

An individual i is well-classified or well categorized as type k if $p_{ik} \simeq 1$. The quality of classification provided by the model is high if all (or most) individuals are well classified. When K increases, we often quickly reach a point at which the p_{ik} values are mostly far away from 1 and 0. Visual inspection, on Figure 2.2 shows that with our model, the quality of classification is good for $K = 3$.

To push the analysis further, we estimated the model for different values of K and looked at different criteria, including entropy, to choose the best model. The difficulty here is that the number of parameters (and time needed for estimation) quickly increases with K (it is already difficult to estimate our model with 4 types). Table B.16 presents the values of different criteria when K varies from 1 to 4.

There exists a tension between Information and Entropy criteria. Celeux and Soromenho (1996) have proposed a choice criterion based on the notion of entropy, called the *Normalized Entropy Criterion*, or NEC. In our context, entropy $\mathscr{E}$ must be defined as follows,

$$\mathscr{E}(K) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{p}_{ik}\ln(\hat{p}_{ik}), \tag{B.11}$$

where $\hat{p}_{ik}$ is the estimated value of the posterior probability p_{ik}. It is easy to check that $\mathscr{E}(1) = 0$ and $0 \leq \mathscr{E}(K) \leq N\ln(K)$, where N is the number of observations i.[15] Entropy is minimal (and equal to zero) when partitioning is perfect.[16] We can divide entropy by its maximum value to obtain an index taking values in $[0,1]$. Define $\mathfrak{E}(K) = (N\ln(K))^{-1}\mathscr{E}(K)$. This index should be minimized.

[13] If q is the number of parameters, N the number of observations and $\mathscr{L}$ is the log-likelihood, then $AIC = 2q - 2\mathscr{L}$ and $BIC = q\ln(N) - 2\mathscr{L}$.

[14] For references on these problems and the discussion of other information criteria, see Celeux and Soromenho (1996).

[15] The entropy is maximal when $p_{itk} = 1/K$ for all k and all i. Entropy is maximal when types cannot be distinguished because any observation can belong to every group with the same probability $1/K$.

[16] Indeed, if for all i, there exists a type $k = k(i)$ such that $p_{ik} = 1$, then, $\mathscr{E}(K) = 0$.

To define the NEC, we consider the gains, in terms of the Log-Likelihood, with respect to $K = 1$, that is $\mathscr{L}(K) - \mathscr{L}(1)$. Entropy is now divided by this gain. NEC is defined as follows,

$$\text{NEC}(K) = \frac{\mathscr{E}(K)}{\mathscr{L}(K) - \mathscr{L}(1)}. \tag{B.12}$$

Another simple criterion that measures the quality of classification is the Average Hirschman-Herfindahl Index. This index is defined as follows,

$$H(K) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{p}_{ik}^2 \tag{B.13}$$

Note that H is equal to 1 if all observations i are perfectly classified. In addition we have, $1/K \leq H(K) \leq 1$. It follows that the lower bound of H is decreasing with K.[17] A normalized index can be constructed as follows. For $K > 1$, define $\mathfrak{H}(K) = (K.H(K) - 1)/(K - 1)$. We have $0 \leq \mathfrak{H}(K) \leq 1$. H and $\mathfrak{H}$ may increase with K ; if these indices drop, this is because the quality of classification deteriorates as K increases. The most important information shown by Table B.16 is that the Log-Likelihood increases markedly until $K = 3$. The marginal gain of adding a fourth type is clearly smaller. So, three types seems a reasonable choice at first glance. A difficulty is that AIC and BIC are always decreasing — they probably reach a minimum for $K > 4$ — but lead to the same conclusion that $K = 3$ is reasonable. The Average Herfindahl and Normalized Herfindahl indices suggest $K = 2$ as the best choice. The normalized entropy $\mathfrak{E}$ clearly indicates $K = 3$, while Celeux and Soromenho's NEC indicates $K = 2$, but NEC doesn't increase much between $K = 2$ and $K = 3$ while it increases a lot more between $K = 3$ and $K = 4$. We therefore choose $K = 3$ as our compromise: not too many parameters, a good classification of individuals and the gains if $K \geq 4$ are apparently small.

B14.2 Estimation of the model by cohorts separately

The model presented above has been estimated with a sample stacking three cohorts of males. The model is very flexible in the sense that most parameters vary by cohort and by type. For this reason, in a nutshell, we find very similar results when the model is estimated with three types on each of the three cohorts separately. Table B.17 very clearly shows that the classification of individuals in three types is very stable to the extent that we find a closely related classification if we estimate the model in a single cohort. Table B.17 gives the correlation matrices of the p_{ik} estimated in the three-cohort model with the estimated p_{ik} obtained in a three-type version of the same model, estimated on a single cohort. The structure of these correlation matrices, with a positive diagonal and high coefficients around .9 and negative off-diagonal values show that our three-type structure does not strongly depend on the fact that we stacked three cohorts. In addition, the full estimation

[17] On the use of H in classification problems, see Windham and Cutler (1992).

results on the three cohorts taken separately do not show big differences.[18] This is reassuring, because one could have suspected that the structure of the economy has changed with time in a manner that our variables do not explain well. Yet, the three-cohort model is very flexible with most coefficients depending on the cohort: this very flexibility probably explains that subsample estimation does not lead to markedly different results.

B15 Online Appendix: Construction of the Sample

In this work we exploit the CEREQ surveys called *Enquêtes Generations à 7 ans*, from 1998, 2004 and 2010. The surveys provide observations during the first 7 years of career of a large representative sample (*i.e.*, cohort) of individuals. The sample includes only individuals who left the educational system during the first survey year (*i.e.*, 1998, 2004 or 2010) and did not return to education during the 7 years of the observation period, except maybe for short on-the-job training sessions. Each of the three stacked surveys contains 3 files: *employment spells, non employment spells and individual characteristics*, the three files form a dataset containing the sequence of employment and unemployment (or non employment) spells for each individual during 7 years

Changes in working hours during employment spells are described. In 1998, the employment-spells dataset contains 47,936 observations, the unemployment dataset contains 30,329 observations and the individuals' file contains 16,040 observations. The corresponding figures are 39,101, 22,724, and 12,365 in the 2004 survey; these figures are respectively 26,056, 16,467, and 8,882 in the 2010 survey.

In each survey, we start by removing the employment spells that are labelled as *family help (i.e., aide familial or afa), self-employed (i.e., à son compte or asc), undescribed summer jobs (i.e., vac)*. This amounts to removing 3,148 employment spells in 1998, 3,572 employment spells in 2004, 2,076 employment spells in 2010. It follows that an individual who is always self-employed (or categorized as *afa*, or *vac*) in the first 7 years after having left the educational system disappears from the data. Then, we merge the employment and non-employment data sets: each individual's history appears with a sequence of employment and non-employment spells. In 1998 we have 75,117 spells, in 2004 58,253 spells, in 2010 40,467 spells.

Individuals are interviewed at the end of their 3rd, 5th and 7th year. They are asked to describe their recent history and their situation at the very moment of the call. So, for each individual, we have 3 additional observations that are the description of their situation at the month of the interview. We recover this information from the 3rd and 5th year of each cohort (*i.e.*, survey) for the individuals observed at the end of the 7th year and we add these data to the 7th year survey. This increases the number of point observations in each cohort, that, at this point are: 29,986 in 1998, 23,011 in 2004, 16,153 in 2010.

[18] The complete cohort-by-cohort results are available upon request.

We deleted the employment spells that lack the working time information; as a consequence, we lose 413 observations in 1998, 66 observations in 2004 and 1,536 observations in 2010.

At this point the beginning and the end of each spell plus the observations at the time of the survey are kept as observations of the individual. Each row of the database becomes an observation (i,t) in the labor market of an individual i (either employed or not), at a date t. At this point the number of observations are : 171,258 in 1998, 133,211 in 2004, 91,174 in 2010.

Each individual enters the dataset the month after the end of his(her) education. There is a date system for each cohort. *Beginning* is the date when an individual in the cohort can be first observed, while *End* is the date of the last observation of the dataset:

- Cohort 1998. Beginning: 1 = January 1998; End: 96=December 2005.
- Cohort 2004. Beginning: 1 = November 2003; End: 98 = December 2011.
- Cohort 2010. Beginning: 1 = November 2009; End: 98 = December 2017.

At this point, the dataset can be described as follows:

- Cohort 1998: 15,950 individuals that are observed on average 10.74 times; (minimum 1, 1st quartile 6; median 10; 3rd quartile 14; maximum 54)
- Cohort 2004: 12,233 individuals that are observed on average 10.89 times; (minimum 1, 1st quartile 6; median 10; 3rd quartile 14; maximum 63)
- Cohort 2010: 8,774 individuals that are observed on average 10.39 times; (minimum 1, 1st quartile 6; median 9; 3rd quartile 13; maximum 45)

Then, we build the experience variable as the sum of working time up to time $t-1$. For each spell we add the information regarding the accumulated experience at time $t-1$ at the beginning, and the end of the spell.

Now, using the individual dataset we create the variables: father is a professional, place of residence at grade 6 entry and the education level (*i.e.*, degree category). The detail for these variables, for each cohort, can be found in the tables below.

The real salary is computed in July 2013 euros.

Then, we remove individuals lacking an observation of the father's occupation and of the residence at grade 6 entry. This leads us to delete 1,071 individuals in the 1998 cohort, 647 individuals in the 2004 cohort and 856 individuals in the 2010 cohort. *Finally, we take the subset of males*. The final dataset for each cohort includes 16,404 individuals, among which:

- Cohort 1998: 80,006 observations for 7,383 individuals;
- Cohort 2004: 60,907 observations for 5,500 individuals;

- Cohort 2010: 37,489 observations for 3,521 individuals.

We stack the three cohorts and generate a unique dataset. We generate a cohort variable c taking values 1998, 2004 or 2010, and a common calendar for the three cohorts where 1 = January 1998 and 240 = December 2017.

Table B.18 lists the degree types that have been aggregated in each of the categories used for estimation.

Table B.18: AGGREGATION OF DEGREES

Education level	Education level detail
	1998 Cohort
Less than High School	SEGPA, reached grades 7 to 11, first year of CAP or BEP, CAP without degree, BEP without degree, CAP, BEP, MC post CAP-BEP, Bac Pro without degree, Brevet or Bac techno without degree, finished grade 12 without degree
High-School Degree	Bac Pro, Bac techno, Bac général, 2 years of College without degree BTS or DUT without degree
Some College, Bachelors	DEUG, BTS DUT, Bac + 3, Bac + 4 IUFM : admitted, IUFM : not admitted
Masters	Bac + 5 and more Excluded: Doctorate and advanced medical degrees
Bus. and Engin. Sch. Deg.	Business Schools, Engineering schools
	2004 Cohort
Less than High School	without degree, CAP, BEP, MC
High-School Degree	Bac pro, Bac techno, Bac général
Some College, Bachelors	Bac+2, DEUG Licence pro, L3, M1
Masters	M2 Humanities, Business adm., Law, M2 Maths, Sciences, Technology, Health, Physical education
Bus. and Engin. Sch. Deg.	Business Schools, Engineering schools
	2010 Cohort
Less than High School	Without degree, CAP, BEP, MC
High-School Degree	Bac Pro, Brevet de Technicien, Brevet Professionnel Bac Techno, Bac général

Continues on the next page...

... table B.18 (continued)

Some College, Bachelors	BTS or DUT other Bac+2 Bac+2/3, Licence pro L3, other Bac+3 M1, Bac+4
Masters	M2 Humanities Business adm. Law M2 Maths Sciences Technology other Bac+5
Bus. and Engin. Sch. Deg.	 Bac+5 Business Schools, Engineering schools

Note. Without degree *i.e., Non-diplômé* means without the diploma or certificate: students who studied but were never granted the degree. Bac is shorthand for baccalauréat (high-school graduation). *Bac pro* means baccalauréat professionnel. *Bac techno* means baccalauréat technologique. Both categories are vocational versions of terminal high-school degrees. CAP and BEP and MC (*i.e.*, mention complémentaire) are pre-bac vocational certificates. *Brevet* is a certificate typically obtained at the end of grade 9. DEUG means two successful years of College. DUT and BTS are vocational degrees, equivalent to the American associate degree. L3 is a Bachelor (three years of College). *Licence pro* is a three-year higher-education vocational degree. The IUFM are preparation schools for primary school-teachers. SEGPA means special education for students with difficulties (grades 6-9).

Table B.19 gives, for each cohorts the definition of the Urban, Peri-Urban and Rural areas used to construct the corresponding indicators. A difficulty comes from the fact that exact definitions changed with the years, but the classification of cities and towns has not changed much.

Table B.20 lists the occupation categories included (and not included) in the definition of the dummy variable called "Father is a professional".

Table B.5: Multinomial logit: Education choice

Type		1	2	3
Below High-School Degree			*Ref.*	
High-School Degree	× 2004 cohort	0.30	0.29	0.59
		(.09)	(0.10)	(.13)
	× 2010 cohort	0.58	0.16	0.98
		(.1)	(.14)	(.15)
	×Father is a professional	0.87	0.67	1.04
		(.10)	(.13)	(16)
	×Peri-urban	0.003	-0.11	0.006
		(.09)	(.10)	(.14)
	× Rural	0.23	-0.31	-0.17
		(.08)	(0.10)	(.12)
	Constant	-0.68	-0.33	-0.75
		(.07)	(.08)	(.09)
Some College and Bachelors	× 2004 cohort	0.18	0.37	0.77
		(.09)	(.10)	(.12)
	× 2010 cohort	0.28	0.46	0.68
		(.1)	(0.12)	(.16)
	×Father is a professional	1.19	1.31	1.77
		(.10)	(.12)	(.15)
	× Peri-urban	-0.12	-0.25	-0.07
		(.09)	(.10)	(.13)
	× Rural	0.05	-0.62	-0.28
		(.08)	(.09)	(.12)
	Constant	-0.65	-0.15	-0.69
		(.07)	(.07)	(.09)
Masters	× 2004 cohort	0.85	1.74	1.79
		(.17)	(.19)	(.19)
	× 2010 cohort	1.29	2.13	2.33
		(.18)	(.20)	(.21)
	×Father is a professional	2.09	1.97	2.16
		(.14)	(.15)	(.18)
	× Peri-urban	-0.25	-0.57	-0.66
		(.17)	(.17)	(.20)
	× Rural	-0.09	-0.73	-0.66
		(.16)	(.16)	(.18)
	Constant	-3.14	-2.90	-2.59
		(.16)	(.18)	(.16)
Bus. and Engin. School Degr.	× 2004 cohort	0.45	0.99	1.07
		(.21)	(.14)	(.19)
	× 2010 cohort	1.56	0.61	1.98
		(.20)	(.18)	(.20)
	× Father is a professional	2.16	2.31	2.60
		(.16)	(.14)	(.18)
	× Peri-urban	-0.02	-0.67	-0.30
		(.18)	(.16)	(.20)
	× Rural	-0.22	-0.85	-0.42
		(.20)	(.15)	(.20)
	Constant	-3.39	-2.04	-2.64
		(.18)	(.12)	(.16)

Table B.6: Ordered Probit: Individual Employment Rate

Type		1	2	3
1998 cohort			*Ref.*	
2004 cohort		-0.11	-0.22	-0.04
		(.02)	(.04)	(.04)
2010 cohort		-0.23	-0.28	0.07
		(.03)	(.04)	(.05)
1998 cohort	High-school Degree	0.08	-0.07	0.15
		(.02)	(.04)	(.03)
	Some College and Bachelors	0.10	-0.05	0.21
		(.02)	(.03)	(.03)
	Masters	0.19	0.19	0.09
		(.06)	(.09)	(.06)
	Bus. Engin. School Degrees	0.22	-0.07	0.33
		(.06)	(.05)	(.06)
	Experience	0.0179	0.0214	0.0176
		(.0004)	(.0008)	(.0006)
2004 cohort	High-school Degree	0.11	0.11	0.15
		(.02)	(.04)	(.04)
	Some College and Bachelors	0.19	0.20	0.35
		(.02)	(.04)	(.04)
	Masters	0.32	0.22	0.32
		(.04)	(.05)	(.05)
	Bus. Engin. School Degree	0.32	0.19	0.38
		(.05)	(.05)	(.06)
	Experience	0.0167	0.0207	0.0190
		(.0004)	(.0006)	(.0007)
2010 cohort	High-school Degree	0.20	0.19	-0.01
		(.03)	(.06)	(.05)
	Some College and Bachelors	0.30	0.32	0.24
		(.03)	(.04)	(.06)
	Masters	0.31	0.18	0.41
		(.04)	(.06)	(.07)
	Bus. Engin. School Degrees	0.93	0.56	0.20
		(.07)	(.08)	(.06)
	Experience	0.0209	0.0266	0.0206
		(.0005)	(.001)	(.0009)
Father is a professional		-0.07	-0.04	0.03
		(.013)	(.02)	(.02)
Peri-urban		0.07	0.04	0.07
		(.013)	(.02)	(.02)
Rural		0.10	0.06	0.13
		(.012)	(.02)	(.02)
Unemployment		-0.11	-0.18	-0.11
		(.006)	(.007)	(0.01)
Cuts	0-0.3	-0.83	-1.734	-0.78
		(.061)	(.066)	(0.09)
	0.3-0.5	-0.79	-1.724	-0.77
		(.061)	(.067)	(0.09)
	0.5-0.6	-0.71	-1.712	-0.72
		(.062)	(.068)	(0.09)
	0.6-0.8	-0.66	-1.707	-0.69
		(.062)	(.069)	(0.09)
	0.8-1	-0.60	-1.693	-0.66
		(.062)	(.070)	(0.09)

Table B.7: Elastic-net regressions of posterior probabilities

	1998			2004			2010			All cohorts		
	Type 1	Type 2	Type 3	Type 1	Type 2	Type 3	Type 1	Type 2	Type 3	Type 1	Type 2	Type 3
1 year late				0,254	-0,114	-0,141						
2+ years late				0,173	-0,094	-0,079						
Repeated a grade							0,035	0,000	-0,237	0,20	-0,04	-0,16
Has not moved	0,011	-0,005	-0,006									
Urban area	-0,002	-0,011	0,013							0,004	-0,009	0,004
Peri-urban										-0,13	0,11	0,02
Rural area	0,127	0,008	-0,134	-0,002	0,023	-0,021				0,02	0,04	-0,06
Father is French										0,01	0,01	-0,01
Foreigner										-0,06	0,00	0,06
Mother is French										0,04	0,02	-0,06
French acquired										-0,04	-0,05	0,09
Foreigner							-0,046	0,000	0,000	-0,01	0,00	0,00
Father : worker										-0,06	0,04	0,03
unemployed							0,090	-0,078	0,000	0,05	0,03	-0,09
retired										-0,05	0,07	-0,03
at home (has worked)							0,090	-0,078	0,000			
at home (never worked)												
training										-0,11	-0,07	0,18
deceased										-0,04	-0,07	0,11
no answer										0,15	-0,01	-0,14
Mother : unemployed	0,018	-0,015	-0,003									
at home (has worked)							0,000	-0,049	0,000			
at home (never worked)							-0,019	0,000	0,000			
no answer				0,127	-0,095	-0,031						
Father : farmer	0,044	-0,016	-0,028							0,27	-0,11	-0,16
Craftsman, business										-0,03	-0,03	0,07
White collar	-0,081	0,020	0,060							-0,03	0,01	0,02
Technician				-0,155	0,036	0,119				-0,09	0,02	0,07
White collar							0,028	0,000	0,000	0,03	-0,01	-0,02
Blue collar							0,000	0,019	0,000	0,00	0,02	-0,02
Does not know										0,05	-0,03	-0,03
Mother : farmer	0,139	-0,053	-0,086							0,16	0,03	-0,19
Craftsman, business	-0,028	-0,008	0,035							-0,09	-0,06	0,15
White collar										-0,08	-0,02	0,09
Technician										0,00	-0,01	0,01
Blue collar										0,07	0,02	-0,09
Does not know				0,137	-0,081	-0,056	0,000	-0,116	0,000	0,11	-0,10	0,00
Auvergne-Rhone-Alpes										-0,03	0,06	-0,03
North (Hauts de France)										0,01	-0,08	0,07
Provence-Alpes-Cote d'Azur				0,000	-0,001	0,001	0,232	-0,148	0,000	0,12	-0,16	0,04
East (Grand Est)	-0,182	0,058	0,124							-0,14	-0,02	0,16
Occitanie	0,095	-0,079	-0,016	0,083	-0,023	-0,060	0,093	0,000	0,000	0,22	-0,09	-0,13
Normandie				-0,025	0,038	-0,014	-0,100	0,000	0,000	0,00	0,00	0,00
Nouvelle-Aquitaine	0,175	-0,066	-0,109	0,028	-0,011	-0,017				0,20	-0,05	-0,14
Centre-Val de Loire										0,00	0,04	-0,03
Bretagne										0,03	0,02	-0,05
Corse										0,24	0,03	-0,27
Pays de la Loire	0,009	0,000	-0,010	0,000	0,004	-0,004	0,000	0,016	0,000	0,08	0,08	-0,16
Paris	-0,125	0,056	0,069				0,000	0,000	0,000	-0,31	-0,02	0,33
Ile-de-France	-0,024	-0,008	0,032	-0,032	-0,022	0,055	-0,031	0,000	0,025	-0,16	-0,02	0,19
Ile-de-France, St Denis (93)										-0,06	0,00	0,06
West Indies, Islands (DOM)							0,152	0,000	0,000	0,23	-0,17	-0,06
Father, graduate							0,000	0,000	0,056			
does not know							0,130	0,000	0,000			
Mother, graduate							0,000	0,000	0,204			
does not know							0,151	0,000	0,000			

Table B.8: DEVALUATION OF DEGREES

	Less than High-school	High-School	Some College Bachelors	Masters (M2)	Bus. and Eng. Schools
2004	0.0663***	0.0477***	0.00272	**-0.0671*****	**-0.0497*****
	(0.00287)	(0.00359)	(0.00390)	(0.0114)	(0.00966)
2010	0.0574***	0.0472***	0.0161***	**-0.0918*****	**-0.0644*****
	(0.00391)	(0.00421)	(0.00460)	(0.0116)	(0.01000)
Constant	7.164***	7.225***	7.388***	7.717***	7.846***
	(0.00167)	(0.00238)	(0.00257)	(0.00961)	(0.00719)
Observations	37868	26659	28835	6389	6261
Individuals	5497	4138	4630	1092	1047

Note. Results obtained by means of OLS on the panel obtained by stacking three 7-year Generation surveys 1998, 2004 and 2010. The dependent variable is the logarithm of the monthly real wages of male individuals with a full-time job. The 1998 cohort is the reference. Stars indicate degrees of statistical significance of the estimated coefficients; * for a p-value¡0.05, ** for a p-value¡0.01 and *** for a p-value¡0.001.

Table B.9: YEARLY RETURNS TO POTENTIAL AND EFFECTIVE EXPERIENCE OF MEN

MEN		Potential Experience		Effective Experience	
$\gamma_k = (1+c_k)^{12} - 1$		OLS	FE	OLS	FE
High School and less	1998	0.0339	0.0372	0.0444	0.0437
Some College and Bachelors		0.0511	0.0533	0.0638	0.0572
Masters and schools		0.0572	0.0564	0.0733	0.0585
High School and less	2004	0.0200	0.0224	0.0320	0.0289
Some College and Bachelors		0.0325	0.0320	0.0421	0.0377
Masters and schools		0.0468	0.0449	0.0665	0.0499
High School and less	2010	0.0237	0.0309	0.0411	0.0403
Some College and Bachelors		0.0393	0.0387	0.0498	0.0431
Masters and schools		0.0449	0.0442	0.0603	0.0477

Note. Results obtained with pooled data stacking the 7-year Generation surveys of 1998, 2004 and 2010, considering males only. The dependent variable is the logarithm of real-wages of individuals with a full-time job. For potential experience as well as for effective experience, the first column on the left gives the OLS estimates, the second column on the right gives the *within*, fixed-effects estimates. Regressions are weighted, using the CEREQ survey weights. All the displayed c_k coefficients are significant at the 1% level.

Table B.10: RETURNS TO DEGREES AT ZERO EXPERIENCE, WITH OR WITHOUT CONTROL FOR THE VARIATION OF UNEMPLOYMENT

Men	1998		2004		2010	
Control for Unemployment	No	Yes	No	Yes	No	Yes
Dropouts	.	.	0.103***	0.106***	0.0552***	0.0620***
	(.)	(.)	(0.0118)	(0.0118)	(0.0140)	(0.0141)
Vocational Degree	0.0212**	0.0216**	0.128***	0.131***	0.127***	0.133***
	(0.00725)	(0.00726)	(0.00818)	(0.00821)	(0.0109)	(0.0110)
High-School Degree	0.0546***	0.0549***	0.150***	0.153***	0.132***	0.138***
	(0.00830)	(0.00830)	(0.00778)	(0.00779)	(0.00891)	(0.00900)
Associate's	0.107***	0.107***	0.215***	0.218***	0.210***	0.216***
	(0.00745)	(0.00746)	(0.00799)	(0.00801)	(0.00862)	(0.00869)
3 years of College (L3)	0.259***	0.259***	0.219***	0.222***	0.262***	0.268***
	(0.0249)	(0.0250)	(0.0139)	(0.0139)	(0.0293)	(0.0293)
4 years of Colleges (M1)	0.271***	0.271***	0.244***	0.247***	0.304***	0.310***
	(0.0159)	(0.0159)	(0.0131)	(0.0131)	(0.0332)	(0.0332)
Master's (M2)	0.488***	0.489***	0.420***	0.422***	0.442***	0.448***
	(0.0201)	(0.0201)	(0.0151)	(0.0151)	(0.0125)	(0.0126)
Business Schools	0.467***	0.467***	0.596***	0.599***	0.495***	0.501***
	(0.0375)	(0.0376)	(0.0397)	(0.0398)	(0.0276)	(0.0275)
Engineering Schools	0.615***	0.615***	0.571***	0.573***	0.587***	0.593***
	(0.0168)	(0.0168)	(0.0153)	(0.0153)	(0.0147)	(0.0147)
Growth of Unemployment		-0.161***				
		(0.0330)				
Constant	7.029***	7.026***				
	(0.00562)	(0.00568)				
Observations	16,2452					

Note: Results obtained by OLS on pooled data stacking three Generation surveys 1998, 2004 and 2010. The dependent variable is the logarithm of the real wages of individuals with a full-time job. The table gives the coefficients and standard deviations of two regressions giving the returns to degrees at zero experience, with degree indicators interacted with cohort and gender, with or without control for the variation of overall unemployment. The 1998 high-school dropouts are the reference group. Potential experience interacted with cohort and degree dummies are introduced in these regressions, but their coefficients are not reported, to lighten the table. Stars indicate the significance of estimated coefficients; * for p-value ¡ 0.1, ** for p-value ¡ 0.05 et *** for p-value ¡ 0.01. Regressions are weighted using Céreq's survey weights.

Table B.11: Proportion of individuals employed in the public sector, by type

	Overall	Type 1	Type 2	Type 3
1998				
Private	.81	.81	.78	.85
Public	.19	.18	.22	.15
2004				
Private	.83	.81	.83	.88
Public	.17	.19	.17	.11
2010				
Private	.80	.78	.78	.85
Public	.20	.22	.22	.15

Table B.12: Distribution of firm size (number of employees) conditional on type k

	Overall	Type 1	Type 2	Type 3
1998				
Small Firms $[1,9]$	.26	.34	.22	.19
Medium Firms $[10,49]$	.27	.31	.24	.24
Large Firms ≥ 50	.47	.36	.54	.57
2004				
Small Firms $[1,9]$	.26	.33	.23	.21
Medium Firms $[10,49]$	.28	.30	.28	.25
Large Firms ≥ 50	.46	.37	.49	.54
2010				
Small Firms $[1,9]$	.25	.32	.21	.20
Medium Firms $[10,49]$	.24	.25	.24	.21
Large Firms ≥ 50	.51	.43	.55	.59

Table B.13: Average number of employment spells, by type

	Overall	Type 1	Type 2	Type 3
1998	2.95	3.12	2.68	3.07
2004	3.15	3.45	2.90	2.98
2010	2.98	3.17	2.69	3.02

Table B.14: Frequency of a professional father conditional on cohort and type

	Type 1	Type 2	Type 3
All cohorts	19.7%	21.1%	23.8%
1998	13.6%	16.8%	19.2%
2004	24.0%	23.6%	26.1%
2010	25.2%	27.1%	30.9%

Table B.15: Probability of reaching an education level given the type, cohort and a professional father (*i.e.*, $z = 1$)

	Conditional on professional father, $p(h\|k,c,z=1)$								
Conditional on cohort ...	1998			2004			2010		
and conditional on type ...	1	2	3	1	2	3	1	2	3
Less than High-school Degree	0.23	0.14	0.13	0.16	0.10	0.08	0.12	0.09	0.06
High-school Degree	0.28	0.19	0.22	0.27	0.15	0.14	0.29	0.16	0.16
Some College and Bachelors	0.35	0.41	0.44	0.36	0.37	0.38	0.28	0.38	0.26
Masters 2	0.07	0.06	0.07	0.14	0.15	0.22	0.14	0.22	0.27
Bus. Engin. School Degrees	0.07	0.20	0.14	0.07	0.22	0.18	0.17	0.14	0.26

Table B.16: Selection Criteria for the Number of Types

Criterion	1 type	2 types	3 types	4 types
Number of parameters	85	158	231	304
Log-Likelihood $\mathscr{L}(K)$	-167,263	-150,893	-143,745	-141,210
$\mathscr{L}(K) - \mathscr{L}(1)$	0	16,370	23,517	26,053
Adj. R^2 of wage regression	.402	.563	.585	.635
AIC	334,696	302,102	287,952	283.028
BIC	335,351	303,319	289,732	285,370
Average Herfindahl (H)	-	0.89	0.84	0.79
Normalized Herfindahl ($\mathfrak{H}$)	-	0.78	0.76	0.72
Entropy $\mathscr{E}$	-	2825	4391	6160
$\mathfrak{E}$	-	0.2484	0.2436	0.2708
NEC	-	0.172	0.186	0.236
Individuals N	16,404	16,404	16,404	16,404

The figures of Table B.16 are derived from EM estimations of the full model with $K = 1,2,3$ and 4 types.

Table B.17: Correlation coefficients of the posterior probabilities of types p_{ik} estimated in a model with three cohorts, with the corresponding probabilities estimated in a model estimated with a single cohort

	Columns: three-cohort model types								
Rows: single-cohort	1998			2004			2010		
model types	1	2	3	1	2	3	1	2	3
1	0.93	−0.54	−0.48	0.73	−0.35	−0.50	0.94	−0.61	−0.47
2	−0.49	0.84	−0.36	−0.40	0.65	−0.25	−0.47	0.90	−0.42
3	−0.53	−0.27	0.91	−0.44	−0.35	0.94	−0.56	−0.28	0.97

Table B.19: AREA OF RESIDENCE AT GRADE 6 ENTRY

Residence area	Residence area, detail
	1988 Cohort
Urban area	municipality belonging to an urban cluster
Peri-urban	municipality belonging to a peri-urban, outer suburban zone
Rural area	Municipalities belonging to a rural-zone labor market Other localities of rural zones Municipality belonging to the periphery of a rural labor market Ultramarine Municipalities (West Indies, etc.) Foreigner, Unknown
	2004 Cohort
Urban area	Urban cluster
Peri-urban	Mono-polarised Municipality
Rural area	Multi-polarised Municipality, Rural space
	2010 Cohort
Urban area	Large urban areas (more than 10 000 jobs), Intermediate urban areas (5 000 to 10 000 jobs)
Peri-urban	Periphery of large and intermediate urban areas
Rural area	Multi-polarized Municipalities in large urban areas, Small clusters (less than 5 000 jobs), Periphery of small clusters, Other Multi-polarized Municipalities, Isolated communes out of the influence of clusters Foreign, Ultramarine communes

Table B.20: OCCUPATION OF THE FATHER

Occupation of the Father	Occupation of the Father, detail
Not a "professional"	Farmer, Craftsman, Storekeeper, Entrepreneur, Technician, Foreman, Salesman, Associate professional, White collar worker, Blue collar worker, unknown
Father is a "professional"	Executive, Engineer, Learned profession, Professor

Note: "Professional" here is a category including the French *professions intellectuelles supérieures*, typically requiring an advanced higher-education degree.

Appendix C

Third Appendix

C1 Grade Distributions

In this section, we display the kernel distribution of grades during the last year of high school for men Danish graduates from *Gymnasiale uddannelser* (upper secondary education programs) during the years 1997 to 2004. In particular, we showcase the kernel density for the final GPA, as well as the grades in mathematics and Danish. Grades in Danish and math comprise only evaluation for level A classes.

Figure C.1: Kernel Density GPA- H.S. Graduates 1997-2004

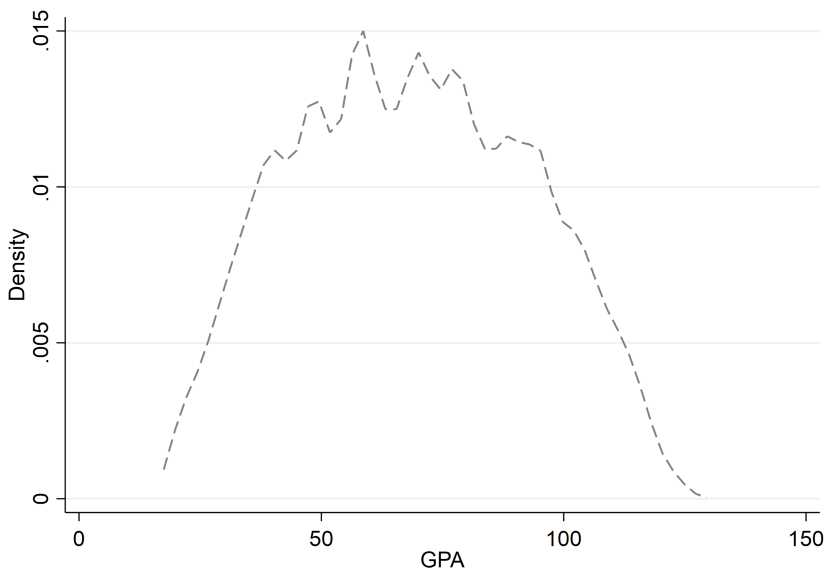
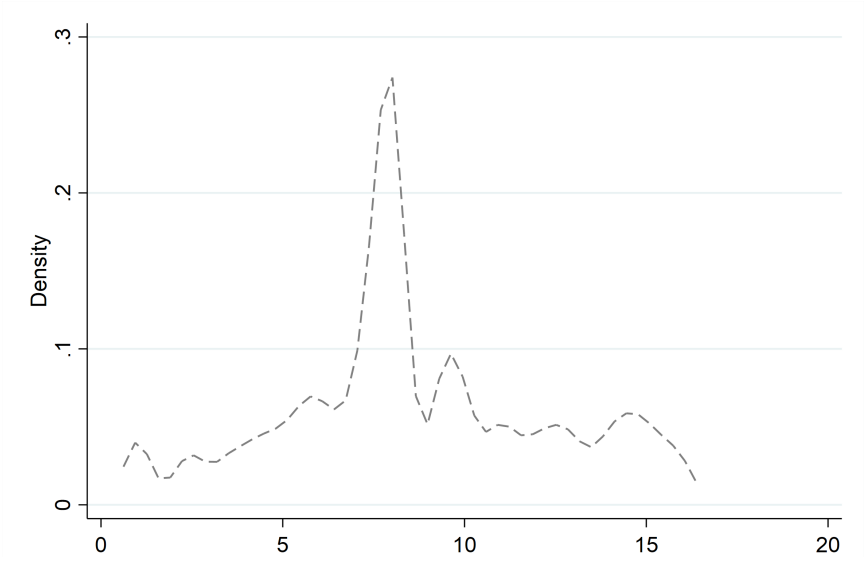


Figure C.2: Kernel Density Math- H.S. Graduates 1997-2004



C2 The Danish HS

The table below provides an overview of the structure of the Danish high school system up to 2004.

C3 GPA and log Wages

In Figure 8, we present the difference in log-wages across different deciles of final H.S. GPA for Danish H.S. graduates from 1997 to 2004, during the observation period spanning 1997 to 2019. Specifically, the displayed coefficients represent the outcomes of a regression analysis, where log-wages are regressed on GPA deciles dummies and year of birth f.e.. Coefficients read as log wage difference with respect to the first decile.

Figure C.3: Kernel Density Danish- H.S. Graduates 1997-2004

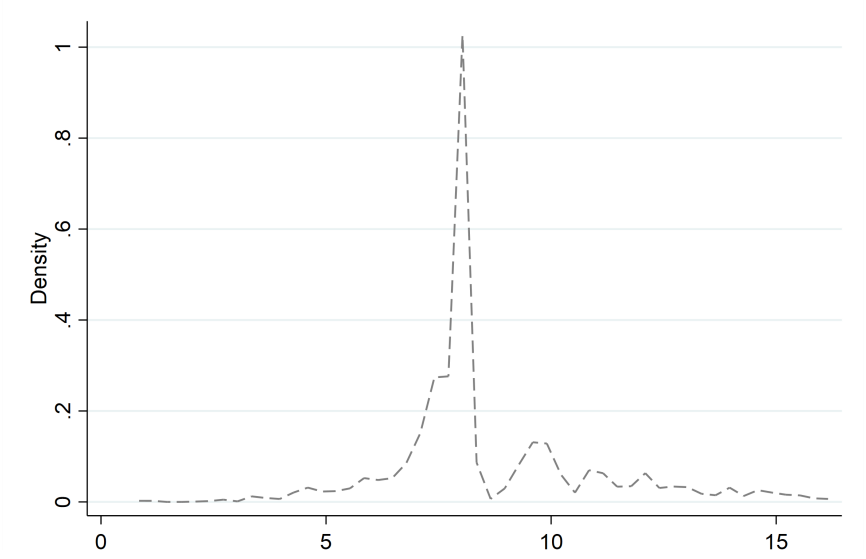


Figure C.4: Caption

General upper secondary education (gymnasium): distribution by subject of the total number of lessons per year (prior to the reform of 2005)

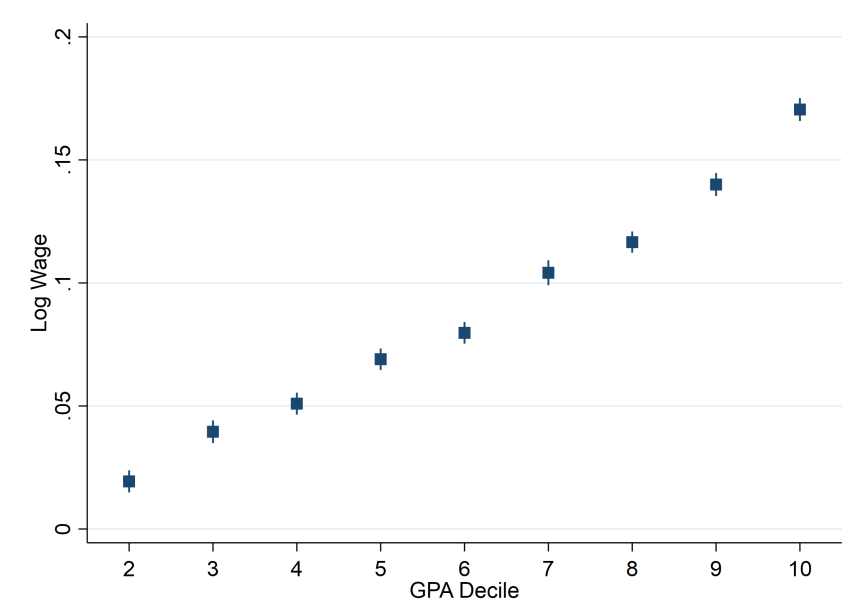
Subject	Number of lessons per year in each form		
	I	II	III
1. *Languages*:			
Beginner language	105	108	–
Visual arts	–	–	51
Biology	79	–	–
Danish language	79	81	102
English	105	108	–
Continuation language	105	108	–
Geography	–	81	–
History and civics	79	81	76
Physical education and sport	53	54	51
Latin	79	–	–
Music	79	–	–
Science	79	108	–
Classical studies	–	–	76
Religious studies	–	–	76
2. *Mathematics*:			
Beginner/continuation language	105	108	–
Visual arts	–	–	51
Biology	79	–	–
Danish language	79	81	102
English	79	108	–
Physics	79	81	–
Geography	–	81	–
History and civics	79	81	76
Physical education and sport	53	54	51
Chemistry	79	–	–
Mathematics	132	135	–
Music	79	–	–
Classical studies	–	–	76
Religious studies	–	–	76
3. *Optional subjects at advanced level*:			
Beginner language	–	–	127
Biology	–	135	127
English	–	–	127
Continuation language	–	–	127
Physics	–	–	127
Greek	–	135	203
Chemistry	–	135	127
Latin	–	135	127
Mathematics, mathematics stream	–	–	127
Mathematics, language stream	–	135	127
Music	–	135	127
Social studies	–	135	127

Source: Danish Eurydice Unit, 2005. Optional subjects are offered at an advanced and intermediate level. Concerning physical education and visual arts as optional subjects, the number of lessons may be combined with the lessons allocated to compulsory subjects. There must be 32 weekly lessons (each lasting 45 minutes) in the first year, and 31-32 in the second and third years. In the first year the 32 lessons are spent on compulsory subjects (27 in the second and 17 lessons in the third year).

Figure C.5: Danish Grading System 7-point grading scale

Grade	Description	ECTS	Old scale (00-13)
12	For an excellent performance displaying a high level of command of all aspects of the relevant material, with no or only a few minor weaknesses	A	13 11
10	For a very good performance displaying a high level of command of most aspects of the relevant material, with only minor weaknesses	B	10
7	For a good performance displaying good command of the relevant material but also some weaknesses	C	9 8
4	For a fair performance displaying some command of the relevant material but also some major weaknesses	D	7
02	For a performance meeting only the minimum requirements for acceptance	E	6
00	For a performance which does not meet the minimum requirements for acceptance	Fx	5 03
-3	For a performance which is unacceptable in all respects	F	00

Figure C.6: GPA Deciles and Log Wages



Notes — Men, H.S. Graduates between 1997 to 2004. Period of observation: 1997-2019. Log wage on GPA decile and year of birth dummies controls.