UNIVERSITÀ
DEGLI STUDI
DI PADOVA

PhD in ECONOMICS AND MANAGEMENT
ECONOMICS CURRICULUM
Department of Economics and Management
University of Padova

Essays in Empirical Economics

Head of the PhD Program: Prof. Antonio Nicolò
Supervisor: Prof. Lorenzo Rocco

PhD Candidate: Elona Harka

Essays in Empirical Economics

Elona Harka

December 2, 2019

Acknowledgements

I would like to express my deepest gratitude to my supervisor Lorenzo Rocco for his excellent supervision, for his continuous and invaluable support, and his patience, his countless hours of guidance, for having taught me how to approach and conduct academic research. I want to thank him in particular for having taught me to pursue my research ideas even when I have encountered difficulties during my research.

I want to express my gratitude also to my co-supervisor, Antonio Nicolò, for his valuable advice on my research and continuous support. My grateful thanks also go to Marco Bertoni, Guglielmo Barone, Fulvio Fontini, Orestis Troumpounis, Giovanni Prarolo and participants of Padova micromeetings.

My thanks also go to Prof. Sascha O. Becker for his hospitality and valuable advice during my visit at the Department of Economics at the University of Warwick. I am also grateful to Helios Herrera and Manuel Bagues.

I am particularly grateful to Hans Goebl for kindly sharing his data for the second chapter of my thesis. I own thanks also to Edgar Haimerl, Laura Vanelli and Simonetta Montemagni.

I thank my parents for their support and encouragement throughout my Ph.D. journey.

Finally, my warmest gratitude goes to my husband Gianluigi. Although he probably had to hear more than he would ever wish about economics and identification strategies, his support has been essential throughout this journey. I couldn't complete it without his unconditional love.

Introduction

This thesis is composed of two essays in Empirical Economics. The first chapter titled "Studying more to vote less. Education and voter turnout in Italy" is co-authored with Lorenzo Rocco. In this study, we use Italian municipality data on education and voter participation in national elections to estimate the effect of schooling on voter turnout. By adopting a fixed effect instrumental variable identification strategy, we find that education reduces voter turnout, more so in municipalities with higher income, lower social capital, which experienced political misconduct in the past and have low institutional quality. Analysis with individual data confirms these results. We discuss several mechanisms to rationalize our findings ranging from the opportunity cost of time to disaffection and civic protest. The second chapter is single-authored, and is titled "Dialects, human capital and labour market outcomes". This paper investigates whether linguistic similarity to Standard Italian affects educational attainment and labor market outcomes, by exploiting unique micro linguistic data on dialects spoken at the beginning of the 20th century in 338 Italian municipalities. Adopting the view that, historically, dialects were the native language of the population and Standard Italian a "second" language, I advance the idea that historical dialects influence contemporary outcomes. Reduced form estimates, exploiting within province variation, show that linguistic similarity to Standard Italian is positively associated with present day educational outcomes and incidence of occupations intensive in language skills.

Contents

1 Studying more to vote less.
Education and voter turnout in Italy. 1

1.1 Introduction . . . 1

1.2 Related Literature . . . 3

1.3 Legislation on compulsory schooling . . . 5

1.4 Data . . . 6

1.5 Empirical strategy . . . 8

1.6 Results . . . 11

1.7 Robustness checks . . . 12

1.8 Heterogeneous Effects . . . 14

1.8.1 Data and definitions . . . 14

1.8.2 Findings . . . 15

1.9 Individual level analysis . . . 18

1.10 Conclusions . . . 19

2 Dialects, human capital and labour market outcomes 43

2.1 Introduction . . . 43

2.2 Related Literature . . . 46

2.3 Background and Linguistic data . . . 49

2.3.1 Background . . . 49

2.3.2 Linguistic Data . . . 51

2.4 The data . . . 52

2.4.1 Contemporary variables . . . 52

2.4.2 Historical Variables . . . 53

2.5 Empirical Strategy . . . 55

2.5.1 Results . . . 57

2.6 Possible mechanisms . . . 61

2.7 Robustness checks . . . 65

2.8 Conclusions . . . 68

List of Tables

1.1 Descriptive Statistics - Municipality level data. 24
1.2 Education and voter turnout 25
1.3 Alternative specifications 26
1.4 Robustness checks - alternative definition of the instrument 27
1.5 Education, urbanization, and north-south dimension 28
1.6 Education, voter turnout, social capital and income 29
1.7 Education, voter turnout, charges of malfeasance and crime 30
1.8 Education, voter turnout, judicial inefficiency, corruption and infrastructure 31
1.9 Education, blank and invalid ballots 32
1.10 Descriptive statistics - Itanes survey 33
1.11 Education and voter turnout - individual level analysis 34
1.12 Description of Variables 40
1.13 First Stage 41
1.14 First stage - Individual level analysis 42

2.1 Linguistic similarity between some dialects 73
2.2 Dialect similarity 73
2.3 Descriptive Statistics of the municipality-level data 74
2.4 Descriptive Statistics of the province-level data 75
2.5 Dialects and years of education 76

2.6 Dialects and upper secondary schooling 77
2.7 Dialects and labour market 78
2.8 Nonlinearities: Effects by quartile of dialect similarity 79
2.9 Dialects and education - LFS individual level data 80
2.10 Dialects and test scores - Invalsi individual level data 81
2.11 Placebo regressions - LFS 82
2.12 Placebo regressions - Invalsi 83
2.13 Mean years of education and linguistic similarity to dialects spoken in big cities 84
2.14 Upper secondary education and linguistic similarity to dialects spoken in big cities 85
2.15 Labour market and linguistic similarity to dialects spoken in big cities 86
2.16 Dialect similarity and literacy 87
2.17 Robustness to omitted variable bias 88
2.18 Robustness checks - range of coefficients sub-samples 88
2.19 Description of contemporary variables 97
2.20 Description of historical variables - municipality level data 98
2.21 Description of historical variables - province level data 99
2.22 Accounting for spatial correlation 101

List of Figures

1.1 Reforms of compulsory education 35

1.2 Voter turnout across Italian municipalities 36

1.3 Voter turnout by cohort and election - Itanes survey 37

1.4 Voter Turnout across municipalities. By year 38

1.5 Mean years of education across municipalities. By year 39

2.1 Localities (Lombardy) 89

2.2 The word “I ragazzi” 89

2.3 Localities 90

2.4 Map of linguistic similarity to Standard Italian 91

2.5 Mean years of education and randomly assigned dialect similarity . 92

2.6 Upper secondary education and randomly assigned dialect similarity 93

2.7 High skilled workers and randomly assigned dialect similarity 94

2.8 Geographical location of municipalities in Table 2.1 95

2.9 Distribution of the index of relative identity value 96

Chapter 1

Studying more to vote less. Education and voter turnout in Italy.

Elona Harka Lorenzo Rocco

1.1 Introduction

Motivated by the classical paradox of voting, social scientists have investigated the reasons for individuals spend time and resources to vote in political elections despite they know that the probability of being pivotal is negligible. As recent evidences point out that a higher voting participation is associated with more egalitarian income distribution (Mueller and Stratmann, 2003), better governance (Glaeser et al., 2007) and larger spending in welfare (Fumagalli and Narciso, 2012), understanding what pushes individuals to vote becomes even more relevant.

Education has been indicated as a positive determinant of voter turnout because more educated individuals are more aware of the issues at stake in an election, more informed, they better understand voting and registration procedures, and they may have attended courses of civics at school (Milligan et al., 2004, Dee, 2004). In this paper, we question this view and we document that in Italy the effect of schooling on voter turnout is actually negative.

While the empirical literature typically relies on individual self-reported voting, our paper takes a different approach by exploiting administrative data on voter turnout at the municipal level in two Italian parliamentary elections held more than 10 years apart. We regress voter turnout defined at the municipal level on the average education of the municipal population. To address the problem of confounders, we add municipality fixed effects and adopt an instrumental variable strategy which relies on a series of reforms of compulsory education, akin in spirit to Milligan et al. (2004). Specifically, we instrument average education in a municipality by the average years of compulsory education assigned by law to the municipality residents. In Italy the length of compulsory education varied over time due to three reforms, enacted since the Sixties, which made compulsory education significantly longer for younger cohorts. Our instrument exploits both the variation in the length of compulsory education across cohorts, the variation in the cohort structure across municipalities and, within municipalities, over time. An analysis with individual-level data from the Italian National Election Studies, which exploits a similar IV strategy, lends support to our central finding that additional schooling adversely affects turnout.

Italy is characterized by marked regional differences in economic development, level of social capital, crime, and quality of local institutions. We explore whether there are heterogeneous effects along these dimensions, and we find that the negative effect of education is stronger in more economically developed areas, but also in areas poorer of social capital, areas which experienced political misconduct, have high levels of crime and poor institutional quality. The latter findings are compatible with a theory according to which the most educated people, who are typically better informed and more aware of the prevailing political practices, abstain from voting as a form of civic protest. The evidence that education is associated with higher proportions of blank and invalid votes supports this conclusion.

The remainder of this paper is organized as follows. Section 1.2 summarizes the relevant literature. Section 1.3 briefly describes reforms on compulsory schooling in Italy. Section 1.4 introduces our data. Section 1.5 presents our empirical strategy. Section 1.6 include the main findings. A battery of robustness checks are discussed in Section 1.7 and the heterogeneity of education effects in Section 1.8. An analysis with individual data is presented in Section 1.9. Conclusion follows.

1.2 Related Literature

The relationship between education and voter participation has been explored both by political scientists and economists. Despite the central role that is commonly attributed to education in enhancing various forms of political participation, literature attempting to establish a causal relationship is rather small and provides contradictory results. Research on the U.S, (such as Dee, 2004; Milligan et al., 2004), finds that education positively affects voter participation. On the other hand, research on European advanced democracies (Milligan et al., 2004; Pelkonen, 2012; Siedler, 2010) does not find any statistically significant effect. In other institutional settings, like authoritarian regimes, Croke et al. (2016) suggest that education adversely affects political participation.

Similarly to us, Milligan et al. (2004) exploit reforms of compulsory education for identification and find a positive relationship between education and the probability of voting in the US. Interestingly, they do not find any significant effect in the UK. Dee (2004) adopts as instrumental variables the availability of junior and community colleges and changes in teen exposure to child labor laws. He confirms that in the US the link between educational achievement and voting participation is positive. To assess the robustness of the above studies, Tenn (2007) takes a different approach, which compares individuals who are about to obtain a given level of education with individuals one year older who have just attained it. The author finds no statistically significant impact of one more year of education on voting participation. This result is also confirmed by Berinsky and Lenz (2011), who exploit the temporal variation in draft procedures and the rates at which men were called to service during the Vietnam War, to assess the causal impact of higher education on political participation. Evidence with experimental data lend support to the existence of a positive link between schooling and voting in the US. For instance, Sondheimer and Green (2010) track the participants of three experiments aimed to increase educational attainment and, throughout a bivariate probit regression model, suggest that high school graduation favors a higher voter turnout.

Turning to Europe, two papers exploit reforms of compulsory education. Siedler (2010) analyzes Germany and Pelkonen (2012) studies Norway and both do not find evidence of a causal link between education and turnout. The latter paper is the closest to ours because the analysis is conducted both at the individual and at the municipality level.

Among developing countries, Parinduri (2016) studies Indonesia and exploits a reform of the school calendar, which increased the length of one particular school year. Also in this case there is no evidence that schooling makes individuals more likely to vote.

Croke et al. (2016) focus on Zimbabwe, an autocratic regime, and conjecture that in such institutional settings, educated citizens intentionally avoid participating in political life. Identification rests on a reform which expanded enrollment in secondary school and results confirm that education decreases the likelihood to vote.

1.3 Legislation on compulsory schooling

Italy passed various reforms of compulsory education in his history. In this paper, we exploit those enacted since the Sixties. In December 1962, the Law 1852 unified the previous lower secondary schools in a unique middle school and made it compulsory starting from October 1, 1963 (Brandolini and Cipollone, 2002). Accordingly, the number of years of compulsory education increased from 5 to 8.[1] This reform has been widely used as a source of exogenous variation in the literature which studies returns of education in Italy (see, for instance, Brunello et al., 2009 among others). Several papers assume that the pivotal cohort is that of 1949. Instead, according to Checchi (2003, p. 3) the reform affected individuals born after 1952. Overall, it is rather difficult to establish which was the first cohort affected by the Law. We take the stand that the pivotal cohort is 1952, i.e. pupils aged 11 years old in 1963, who after primary education were eligible to enroll in middle school in October 1963. Two reasons justify this choice. First, individuals of earlier cohorts, who completed primary school between 1960 and 1962, but were below age 15 in 1963, were free to abandon education at the end of primary school, and many did so. Second, according to the art.5 of the Law 1852/1963 individuals aged 12, 13 and 14 in 1963 (cohorts 1951, 1950 and 1949) could be admitted to the second grade, the third grade, or the final exam respectively of the middle school, upon request. This norm implies that completing middle school was an option and not an obligation.

Years of compulsory education increased again in 1999 from 8 to 9. More specifically, students were allowed to quit after completing the second grade of the upper secondary, or at age 15, with at least nine years of schooling. The Law was ab-

[1] Retained pupils were allowed to drop out at age 15 after at least 8 years of schooling.

rogated in 2003 and compulsory education reverted to 8 years. Accordingly, the reform of 1999 affected only few cohorts. In our empirical strategy, we follow Brilli and Tonello (2014) and Vergolini and Raimondi (2017), who suggest that the affected cohorts were those born between 1985 and 1988.

Three years later, the Law 296/2006 extended compulsory education again, this time to 10 years, and also increased the minimum leaving age to 16 years, starting from the academic year 2007/2008. In this case, the pivotal cohort is that born in 1993, corresponding to students aged 14 in 2007. Figure 1.1 summarizes how these reforms changed the length of compulsory education across cohorts.

1.4 Data

We use data on voter turnout in the national parliamentary elections held in 2001 and 2013 by municipality, matched with census data from 2001 and 2011, which provide detailed information on education and the age structure of the resident population, by municipality.[2] In general elections, Italian citizens vote for both the Chamber of Deputies (lower house) and the Senate of the Republic (upper house). The minimum voting age is 18 for the lower house and 25 for the upper house. In parliamentary elections only Italian citizens have the right to vote, while immigrants who reside in Italy, but who are not citizens, are excluded. Voting by post is not allowed[3] and voters who live in a municipality other than that of legal residence are required to move into the latter to cast their ballot. The electoral rule was majority voting with a proportional correction in 2001 and proportional voting with a majority premium in 2013. Both electoral rules induced parties to form large coalitions and can be roughly assimilated to majority voting.[4]

[2] Parliamentary elections were also held in 2006 and 2008. We focus on the elections which are closest to the census years.

[3] The only exception are Italian citizens living abroad who vote for reserved parliamentary seats.

[4] Aosta Valley and Trentino Alto Adige regions, which host large linguistic minorities, adopted different rules and therefore are excluded from the analysis.

We focus on the lower house and voter turnout is defined as the share of voters out of all the eligible voters. Data are drawn from the Archivio Storico delle Elezioni of the Ministry of Interior.

Italy has traditionally had high levels of voter turnout, although declining since the Nineties. In our sample, the average turnout declined from 80.9 percent to 75.7 percent between 2001 and 2013. Figure 1.4 provides a visual description of voter turnout across municipalities and over time.

Since concurrent local elections might alter voter turnout, we identify municipalities in our sample who held local election the same day of the national elections. In 2001, 963 municipalities had municipal elections, while in 2013 the regions of Lombardy, Molise and Lazio had regional elections.

Counts of the population in voting age by level of education and municipality are not publicly available and have kindly been provided by ISTAT. From these data we compute the average years of education of the population aged 18+ by municipality and year.[5] Unfortunately, such counts are missing when only 3 or less residents attained a given education level.[6] To avoid measurement errors, which are a concern especially for the many small municipalities in the sample, we keep only those municipalities for which we have data for all levels of educational attainment. This choice is not without cost as it reduces our sample from about 8000 to 5861 municipalities.

Average education increased significantly from 2001 (average: 8.4 years) to 2011 (average: 9.4 years), as shown in Figure 1.5. Several reasons explain this variation: 1) older cohorts with lower education were replaced by younger cohorts who are

[5] Average education refers to the residing population while voter turnout is calculated over the pool of citizens which excludes most immigrants. Although immigration in Italy is lower than elsewhere, it increased from 2.3 percent in 2001 to 7.4 percent in 2013 and it is much higher in the more developed North and Centre than in the South. Generally, immigration in Italy is low skilled (Bratti and Conti, 2014) so that average education in the residing population is lower than among citizens.

[6] The reported levels of education are: less than primary, primary, lower secondary, short upper secondary programs of 2 or 3 years, long upper secondary programs of 4 or 5 years, three-year tertiary degree, and 4 or 5-year bachelor.

better educated; 2) the size of old and new cohorts is different; 3) a significant proportion of individuals aged between 18 and 25 in 2001 acquired additional education by 2011; 4) mortality varies by cohort, education, and municipality; 5) a small percentage of adult individuals acquired further education between 2001 and 2011; 6) the pattern of selective internal migration, whereby the more educated tend to flow towards more economically advanced areas, may have changed over time.

We also define an alternative measure of education, based on census data, namely the share of individuals aged 19 and over with at least upper secondary education[7] Table 1.12 in appendix presents in detail data sources and the definition of the main variables. Table 1.1 provides the main descriptive statistics for the data at use, by year.

1.5 Empirical strategy

To estimate the impact of schooling on voter turnout, we employ a fixed effect instrumental variable (FE-IV) strategy. Our baseline specification is:

$$T_{mt} = \theta_m + \gamma_{pt} + \alpha_1 Edu_{mt} + \alpha_2 L_{mt} + X_{mt}\beta + \varepsilon_{mt} \tag{1.1}$$

where T_{mt} is voter turnout at municipality m at time t, Edu_{mt} is average education (either mean years of education or the proportion of individuals with at least an upper secondary degree), and θ_m are municipality fixed effects. Province by time fixed effects are denoted by γ_{pt}, and X_{mt} are municipality-by-time controls, namely the share of females and the presence of concurrent elections. Variable L_{mt} is the proportion of individuals in working age (18-64) and serves to control

[7] Age 19 is the age at which students are expected to conclude upper secondary school.

for the time-varying age structure of the municipal population.[8] Province-by-time fixed effects play an important role in our analysis. First, they flexibly account for the change in electoral rules occurred between 2001 and 2013, which might have altered voter turnout depending on the level of political competition at the local level. Second, they capture the effect of the Great Recession, which was particularly acute in Italy since 2009, and hit stronger the industrial areas. The Great Recession was largely responsible for the rising discontent among Italians and for the support to the anti-establishment Five-Stars Movement, born in 2009. Third, they account for the differential increase in the share of immigrants, much higher in the more economically dynamic North and the Centre than in the South, which has been shown to affect both voter turnout and voting for the extreme right (Barone et al., 2016).[9] Despite all controls, the key explanatory variable Edu_{mt} can be endogenous if there exist time-varying unobservables at the municipality level, which affect both education and voting. For instance, average cognition may change over time within the same municipality for purely random reasons. Alternatively, shocks on the labour market might have induced more residents to acquire additional education and can have simultaneously altered voting participation. A localized industrial downturn, with mass layoffs and rising unemployment, may induce the younger cohorts to stay longer at school, waiting for better conditions on the labour market and, simultaneously, boost political participation and turnout as a mean of protest or to obtain an intervention from the government (Incantalupo (2015), Aytaç and Stokes (2019, p. 107)). In this case the OLS bias would be positive and large.[10] To address this concern, we instrument Edu_{mt} by

[8] When schooling is measured by the share of individuals aged 19+ with at least an upper secondary degree, all controls refer to the population aged 19 and over.

[9] To address the problem that average education is computed over the residing population while voter turnout is defined over the pool of citizens, we experimented a specification which includes the share of immigrants at the municipal level among the controls, despite this is likely an endogenous variable. Reassuringly, results did not qualitatively change.

[10] A small increase in municipal average education, due to the educational choice of relatively few young cohorts, is associated with large variations in turnout.

$MYCS_{mt}$, the average years of compulsory schooling of the population aged 18 and over in municipality m at time t. Formally:

$$MYCS_{mt} = \sum_{a=18}^{100} CS_{at}\pi_{amt}$$

$$\sum_{a=18}^{100} \pi_{amt} = 1 \text{ for each } t = 2001, 2011$$

where CS_{at} is the number of compulsory years of education for individuals of age a at time t, and π_{amt} is the proportion of municipal population which is aged a at time t, for $a = 18...100$. Jointly the vector of proportions π_{amt} represents the age structure of the municipal population.[11] The latter varies across municipalities, because of differences in fertility and mortality, and over time because the age structure of the population changes as time passes. For instance, it is well know that in most Italian municipalities population ages, implying that the distribution by age of the population gets more right-skewed. Also CS_{at} varies over time, because age a in 2001 and in 2011 corresponds to two cohorts 10 years apart. For instance, individuals aged 55 in 2001 belong to cohort 1946 and were expected to stay at school at least 5 years, while individuals aged 55 in 2011 belong to cohort 1956 and had acquired education for at least 8 years. Similarly, compulsory education of individuals aged 18 in 2001 (cohort 1983) was 8 years, while that of their peers in 2011 (cohort 1993) was 10 (see Figure 1.1). While the parametrization in terms of age is convenient and it is the one we adopted throughout the paper because the support of the age distribution remains constant over time, the link between reforms of compulsory education and the instrument can also be appreciated if one reasons in terms of cohorts. At a given point in time, the cohort structure of the population varies across municipalities due to differences in past demographic trends. The cohort structure also varies over time for two reasons:

[11] To be consistent, the instrument is defined over the population aged 19+ when the alternative measure of education is analyzed.

1) older cohorts disappear while new cohorts enter; 2) relative cohort sizes change over time as older cohorts shrink due to higher mortality. This implies that the effect of each reform on the average years of compulsory education varies across municipalities, depending on the size of the cohorts affected by the reform, and over time, due to variations in the cohort structure of the population.

Two final remarks are worth making. First, given the inclusion of municipality and time fixed effect, it is the municipality-by-time component of the instrument which identifies our IV procedure. Second, in model (1) we parsimoniously control for the age structure of the municipal population by including L_{mt}. This is to neutralize the possible correlation, within-municipality, between instrument and age structure. Specifications which control more finely for age structure are discussed later, but we anticipate that results are unaffected.

1.6 Results

For reference, we start by estimating model (1) with municipality fixed effects without instrumenting. In Table 1.2, columns (1) and (3) report results for both measures of municipal education. Fixed effect estimates are positive, but small and only marginally significant at the conventional levels. In particular, results in column (1) suggest that one additional year of average education increases voter turnout by 1.1 percentage points.

The corresponding IV estimates reported in column (2) and (4) contradict these findings. One additional year of average education reduces voter turnout by 7.2 percentage points, while 10 additional percentage points in the proportion of residents with high education reduce voter turnout by 5.3 percentage points.[12]

In interpreting these results we remark two points. First, FE and IV estimates

[12] Alternatively, one standard deviation increase in the average years of education decreases voter turnout by 5.7 percentage points, and one standard deviation increase in the share of individuals with at least an upper secondary degree decreases voter turnout by around 4.2 percentage points.

should be compared with care. While the former has the nature of an ATT (average treatment effect on the treated), the interpretation of the latter is that of a LATE effect (local average treatment effect), which captures the effect of education among compliers. We shall return on this point later.

Second, a unitary increase in mean years of education does not (typically) correspond to a uniform unitary increase among all residents. Rather, in most cases, it corresponds to situations where the variation in individual education can be very heterogeneous. Hence the estimates derived from our municipality-level analysis can be compared with the estimates at the individual level discussed in the literature (and below in Section 1.9) only under the strong hypothesis that the relationship between education and voting turnout is linear over the entire support of education.

Turning to the instrument, the reported F statistics suggest that it is not weak. The large figures (162.9 and 139.7) partly reflect the fact that both the endogenous variable and the instrument embodies the municipality age structure. When we weaken this tie (see the following Section 1.7) the instrument remains strong with F statistics well above 10.

1.7 Robustness checks

In this section we consider four robustness checks. First, we replace L_{mt}, the share of working-age population, with the shares of the municipal population in 12 age bins, from 18-25 to 75 and over, average age[13] and the interaction between average age and a time dummy. Formally we estimate the following model:

$$T_{mt} = \theta_m + \gamma_{pt} + \alpha_1 Edu_{mt} + \sum_{b=1}^{12} \delta_b \pi_{bmt} + X_{mt}\beta + \varepsilon_{mt} \tag{1.2}$$

[13] Computed over the population aged 18+

Where π_{bmt} is the proportion of individuals in the age bin b in the municipality m at time t. In a more detailed specification we replace L_{mt} with the full vector of shares $\pi_{18mt}...\pi_{100mt}$. Estimates are reported in columns (1) and (2) of Table 1.3 respectively, and are in line with baseline results, although standard errors get larger, likely because the additional controls increase the level of model multicollinearity. However, it is reassuring that the F-statistic remains largely above the rule of thumb of 10 also in the most fine-grained specification.

Second, to better account for local shocks, we replace province-by-time fixed effects with local labor market-by-time fixed effects. The results of this specification are reported in column (3) of the Table 1.3 and confirm our central finding.

Third, we address the concern that the age structure of the population used to build the instrument might be correlated with shocks at the municipal level, by using the age structure prevailing in each municipality 5 and 10 years earlier. Formally, instrument $MYCS_{mt}$ is now defined as:

$$MYCS_{mt} = \sum_{a=18}^{100} CS_{at}\pi_{am,t-k}$$

$$\sum_{a=18}^{100} \pi_{amt-k} = 1 \text{ for } t = 2001, 2011 \text{ and } k = 5 \text{ or } k = 10$$

In both cases, the F test suggests that the instrument is not weak, and our basic findings are confirmed and actually reinforced (Table 1.4).

Fourth and last, to ensure that our estimates are not driven by any province in particular, we estimate the baseline model by excluding one province at the time. The results are robust, statistically significant at 1 percent, and the point estimate range from [-0.079, -0.063]. Results are robust also when we omit one electoral district at the time. In this case, point estimates range from [-0.084, -0.0597].

1.8 Heterogeneous Effects

Italy is characterized by sharp regional differences in income levels, quality of local institutions, and social capital, among others. Disparities are observed also in voter participation. Figure 1.2 shows voter turnout across Italian municipalities in the parliamentary elections of 2001. Centre-North municipalities have higher levels of voter turnout. In our sample, mean voter turnout of municipalities in the Centre-North is around 85 percent in the elections of 2001 and 79 percent in 2013, against 73 percent and 67 percent respectively in the Southern municipalities.

To shed light, on how the relationship between schooling and participation in elections is affected by local characteristics, we extend model (1) by including, in turn, interaction terms between education and a specific pre-determined and time-invariant characteristic, namely social capital, income per capita, political malfeasance, crime, and quality of local institutions.[14]

1.8.1 Data and definitions

To explore heterogeneity, we use indicators drawn from various sources. As regards social capital, the literature proposes a variety of proxies. Among them, our preferred are generalized trust and blood donations. Generalized trust is constructed from waves 2 and 3 of the European Values Study conducted in the 1990s (EVS, 2015). Generalized trust is defined as the fraction of the individuals answering "Most people can be trusted", by region. Blood donations correspond to the number of blood bags collected by AVIS per million inhabitants in 1995, by province. Data are from Guiso et al. (2004). Charges of malfeasance refer to the fraction of deputies who received at least one request to remove parliamentary immunity (RAP) between 1948 and 1994 (during the legislatures from I to XI),

[14] The interaction between education and each local characteristic is instrumented by the interaction between $MYCS$ and the same characteristic. Moreover, we include among the controls interaction terms between the local characteristic and the share of the population in the working-age to increase model flexibility.

by electoral district. We derive it from the dataset of Golden (2007).

Crime level is an index derived from a principal component analysis, including the mafia index from Calderoni (2011) and crime rates against the person, the property, and the State, and other crimes per 100,000 inhabitants in year 2000, by province (from ISTAT).

As a proxy for the quality of local institutions, we employ three measures - judicial inefficiency, corruption, and public good provision. Judicial inefficiency data are from Guiso et al. (2004) and refer to the number of years necessary to complete a trial. Corruption is measured through the Golden Picci index (Golden and Picci, 2005), who propose a measure of corruption based "on the difference between the amounts of physically existing public infrastructure and the amounts of money cumulatively allocated by government to create these public works." The larger this difference, the larger corruption. Finally, public good provision refers to the endowment of economic infrastructures of the Italian provinces in 1997 measured by Ecoter (2000). We normalize all these variables in the unitary interval, where 0 and 1 corresponds, respectively, to the minimum and the maximum observed value.

1.8.2 Findings

To begin with, we explore possible heterogeneous effects between rural and urban areas and Centre-North vs. South regions. In column (1) of Table 1.5, we report that one additional year of education decreases voter turnout by 6.4 percentage points in urban areas and by 8.2 percentage points in rural areas. Moreover, the marginal effect of education is negative, but small and not statistically significant in the Centre-North regions, and as high as 16 percentage points in South regions (column (2) of Table 1.5).

Turning to social capital (Table 1.6), the effect of education decreases in absolute value as social capital increases. In particular where the level of generalized trust is high, the effect of education disappears (column (1) of Table 1.6), and surprisingly, it gets even positive in municipalities where blood donations are more frequent (column (2) of Table 1.6). Column (3) of the Table 1.6 shows that the marginal effect of education is larger in absolute value in richer municipalities and ranges between -5.4 percentage points in areas with low GDP per capita to -8.7 percentage points in richer areas.

Charges of wrongdoing reinforce the negative effect of education. In particular, as displayed in column (1) of Table 1.7, when the fraction of deputies subject to a criminal investigation reaches its maximum, the effect of education turns to be as large as -19.8 percentage points. Similarly, education discourages voting more strongly in areas with high criminality (column (2) of Table 1.7).

The quality of institutions also modifies the effect of education. We find consistent results across the three indicators we consider: the effect of education is stronger (more negative) in municipalities where the judicial system is more inefficient, where corruption is higher and where infrastructures are underdeveloped (Table 1.8).

Overall, these results suggest several possible explanations for the negative effect of education on voter turnout. On the one hand, in areas where civism is stronger, education counts less for voter turnout. Citizens are pro-social, are engaged in social activities and political participation is widespread, regardless of education. Likely, in these areas, students even learn civism at school. On the other hand, rather unsurprisingly, where returns to education are higher, and so the opportunity cost of time is larger (that is, in areas with higher GDP per capita), the negative effect of education is stronger.

The modifying effects on misbehavior, crime, and institution quality suggest another explanation, that more educated individuals, the ones who are more aware and informed, complain against the political class by not showing up at the polls, a sort of civic protest. They are disaffected with politics, understand that politics is responsible for the poor performance of institutions, and do not want to legitimize the current political class by participating to elections.

We find some indirect support to this conjecture. We test whether there is an effect of education on the proportion of blank ballots and the sum between blank and invalid ballots.[15] Table 1.9 shows that, as expected, both outcomes increase by respectively 3.3 and 3.4 percentage points for a unitary increase in voters' average education. Relative to an average proportion of blank and invalid votes of about 6.5 percent in the sample, these effects are quite large.

On the other hand, our results do not support an alternative explanation based on internal selective migration. Suppose that many well educated individuals move to more developed areas, which offer better labour opportunities, without simultaneously changing their legal residence. To them, voting would be costly and time consuming, because they have to return to the municipality where they grew up to be able to cast their vote. Then, in the more disadvantaged areas, we should observe a lower (not higher, as we find) turnout among the more educated who are listed in the official registry, because many of them do not actually live there and do not come back to vote.

We mentioned that our IV estimates have the nature of LATE effects, that is of the average effect among the subset of compliers.[16] In our context, compliers are those individuals who acquired more education only because compulsory education expanded, and would not otherwise. In particular, the compliers affected by the reform of 1963, took three additional years of schooling, they attended one hour of civics a week[17] and added relatively advanced notions of Italian literature, maths and history to the basic skills learned at primary school. These notions

[15] Invalid ballots are often too few to allow for a separate analysis.

[16] There is evidence that, particularly in the South, the reform of 1963 took years to be fully implemented. Hence, we cannot exclude the existence of a group of "never takers".

[17] Introduced since 1958 in all lower and upper secondary schools.

might have significantly reinforced these citizens' ability to elaborate information, perceive, understand and, eventually, feel anger for politicians' misconduct.

1.9 Individual level analysis

To make our analysis comparable with earlier research, we turn to individual self-reported data, despite concerns about misreported voting behavior. We rely on the Italian National Election Studies (Itanes), a survey designed as a repeated cross-section, which is carried out soon after any general election, and which includes information on individuals' voting behavior, demographics, educational attainment, and a few other controls. We use data collected after the parliamentary elections of 2001, 2006 and 2008 [18] leaving out the wave collected after the election of 2013 which, according to Itanes itself, is nonrepresentative of the population, especially as regards education and age. Our working sample counts 9322 individuals. Table 1.10 displays key summary statistics and shows, interestingly, that voter turnout is 92.9 percent, well above the levels of voter participation in Italy in the same period, which, according to official data, were about 82 percent. This fact alone confirms our suspicion that self-reported data suffer from sizable measurement error and that individuals tend to over-report vote participation, likely to please the interviewer.

Interestingly, Figure 1.3 suggests that self-reported turnout varies little by cohort, both in 2001 and 2008, and we do not observe marked negative cohort trends, as it is the case for instance in the UK, where the 18-24 vote systematically less than more senior citizens (Pickard, 2019).

We define a linear probability model defined as:

$$T_{it} = \alpha_0 + \theta_r + \gamma_{rt} + \alpha_1 Edu_{it} + X_{it}\beta + \varepsilon_{it} \tag{1.3}$$

[18] The electoral rule for the elections of 2006 and 2008 was the same as in 2013, proportional voting with a majority premium.

where T_{it} is a dummy variable equal to 1 if individual i at time t declares of having voted in the national election and zero otherwise. The main explanatory variable Edu_{it} is years of education, θ_r are region dummies, γ_{rt} are region by time fixed effects and X_{it} are individuals specific controls, including gender and a quadratic polynomial in the year of birth, which accounts for possible cohort-effects. We instrument Edu_{it} by the years of compulsory schooling to which individual i was subjected, depending on his or her cohort. Specifically, instrument variation depends on the reforms of 1963 and 1999.

Table 1.11 shows that the effect of education is positive and very small when estimated by OLS and negative when estimated by IV. These results confirm our findings based on municipality data, although they are not fully comparable as we have remarked above.

1.10 Conclusions

We have analyzed the effect of education on voter turnout in Italy, a country with low barriers to voting participation, but characterized by a marked political instability, frequent general elections, and sustained political malfeasance in the recent past. We have exploited administrative data at the municipality level, free of measurement error, and have relied on a series of reforms of compulsory education to identify causal effects. We have found a sizable negative effect of education on voter turnout. A one year increase in average municipal education causes voter turnout to decline by 7.2 percentage points. The result passes several robustness tests and is confirmed when the analysis is conducted on self-reported individual data.

With the exception of autocratic countries, our findings contrast with the positive effects of education documented in the US and the absence of effect in Northern Europe, despite the identification strategies are comparable. Such difference does not seem due to possible measurement errors, which might plague self-reported individual data, because we do find a negative effect of education in Italy even when we replicate the analysis on survey data. We believe instead that our results depend on the specific Italian context, which was rather peculiar during the study period between 2001 and 2013. In Italy there were, and to a large extent there are, widespread negative feelings and mistrust towards politics and politicians, after the scandal of corruption uncovered by the famous "Many Pulite" judicial investigation in the Nineties.

To shed light on the possible mechanisms, we have explored several dimensions of heterogeneity. We have found that the negative effect is stronger in South municipalities and rural areas, weaker in areas rich of social capital, but stronger again in the more economically developed localities. Moreover, the negative effect of education is reinforced by past records of political misconduct, high levels of crime and poor institutional quality. These results support the hypothesis that more educated people, who are more informed and more aware of the political practice, choose to abstain from voting to express their discontent and protest. By abstaining, they refuse to legitimize the political class. This conclusion is credited by the result that a unitary increase in average education causes the proportion of blank and invalid votes to increase by 3.4 percentage points, which corresponds to an increase of over 50 percent from the sample mean.

The implication of these findings is worrisome. If the more educated withdraw from political participation, contents will be replaced cheap talk in the political debate, and the democratic institutions might be captured by populisms and easy propaganda.

References

Aytaç, S. E. and Stokes, S. C. (2019). *Why Bother?: Rethinking Participation in Elections and Protests*. Cambridge University Press.

Barone, G., D'Ignazio, A., de Blasio, G., and Naticchioni, P. (2016). Mr. rossi, mr. hu and politics. the role of immigration in shaping natives' voting behavior. *Journal of Public Economics*, 136:1 – 13.

Berinsky, A. J. and Lenz, G. S. (2011). Education and political participation: Exploring the causal link. *Political Behavior*, 33(3):357–373.

Brandolini, A. and Cipollone, P. (2002). Return to education in italy: 1992-1997.

Bratti, M. and Conti, C. (2014). The effect of (mostly unskilled) immigration on the innovation of italian regions.

Brilli, Y. and Tonello, M. (2014). Rethinking the crime reducing effect of education: the role of social capital and organized crime.

Brunello, G., Fort, M., and Weber, G. (2009). Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe. *The Economic Journal*, 119(536):516–539.

Calderoni, F. (2011). *The Mafia Index. A measure of the presence of the mafia across Italian provinces*, pages 141–162.

Checchi, D. (2003). The italian educational system: family background and social stratification.

Croke, K., Grossman, G., Larreguy, H. A., and Marshall, J. (2016). Deliberate disengagement: How education can decrease political participation in electoral authoritarian regimes. *American Political Science Review*, 110(3):579–600.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics*, 88(9):1697 – 1720.

Ecoter (2000). La dotazione infrastrutturale nelle province italiane. aggiornamento al 1997.

EVS (2015). European values study longitudinal data file 1981-2008 (evs 1981-2008).

Fumagalli, E. and Narciso, G. (2012). Political institutions, voter turnout, and policy outcomes. *European Journal of Political Economy*, 28(2):162 – 173.

Glaeser, E. L., Ponzetto, G. A., and Shleifer, A. (2007). Why does democracy need education? *Journal of economic growth*, 12(2):77–99.

Golden, M. (2007). goldenpm.zip. In *Charges of Malfeasance, Preference Votes, Government Portfolios, and Characteristics of Legislators, Chamber of Deputies, Republic of Italy, Legislatures I-XI (1948-1994): Parliamentary Malfeasance*. Harvard Dataverse.

Golden, M. A. and Picci, L. (2005). Proposal for a new measure of corruption, illustrated with italian data. *Economics & Politics*, 17(1):37–75.

Guiso, L., Sapienza, P., and Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94(3):526–556.

Incantalupo, M. B. (2015). The effects of unemployment on voter turnout in us national elections.

Itanes (2001). Surveys on voting behaviour in italy. Available at http://www.itanes.org/dati/.

Itanes (2006). Surveys on voting behaviour in italy. Available at http://www.itanes.org/dati/.

Itanes (2008). Surveys on voting behaviour in italy. http://www.itanes.org/dati/.

Milligan, K., Moretti, E., and Oreopoulos, P. (2004). Does education improve citizenship? evidence from the united states and the united kingdom. *Journal of Public Economics*, 88(9):1667 – 1695.

Mueller, D. C. and Stratmann, T. (2003). The economic effects of democratic participation. *Journal of Public Economics*, 87(9):2129 – 2155.

Parinduri, R. (2016). Does education increase political participation? evidence from indonesia.

Pelkonen, P. (2012). Length of compulsory education and voter turnout—evidence from a staged reform. *Public Choice*, 150(1):51–75.

Pickard, S. (2019). *Young People, Voter Registration, Voting, Elections and Referendums*, pages 235–271. Palgrave Macmillan UK, London.

Siedler, T. (2010). Schooling and citizenship in a young democracy: Evidence from postwar germany. *The Scandinavian Journal of Economics*, 112(2):315–338.

Sondheimer, R. M. and Green, D. P. (2010). Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science*, 54(1):174–189.

Tenn, S. (2007). The effect of education on voter turnout. *Political Analysis*, 15(4):446–464.

Vergolini, L. and Raimondi, E. (2017). 'everyone in school': The effects of compulsory schooling age on drop-out and completion rates.

Tables

Table 1.1: Descriptive Statistics - Municipality level data.

	2001		2011(2013)	
	mean	sd	mean	sd
Voter turnout	0.809	0.0974	0.757	0.0709
Mean years of education	8.412	0.621	9.399	0.624
Upper secondary education	0.271	0.0645	0.357	0.0686
Share of female	0.528	0.0312	0.534	0.0294
Concurrent elections	0.164	0.371	0.270	0.444
Working-age population (18-64)	0.763	0.0553	0.744	0.0483
Observations	5861		5861	

Note: Municipality level data. *Voter turnout* is defined as the share of eligible voters who cast a ballot in parliamentary elections of 2001 and 2013 (source: Ministry of Interior). *Concurrent elections* is a dummy which takes 1 if the municipality voted for a local election the same day of national elections (source: Ministry of Interior). *Mean years of education* is the average years of schooling in the population aged 18+.*Upper secondary education* is the proportion of residents with at least an upper secondary degree over the population aged 19+. *Share of females* is the proportion of females aged 18+. *Working-age population* is the share of the population aged 18-64. These variables are based on 2001 and 2011 census data provided by ISTAT.

Table 1.2: Education and voter turnout

	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
dep. var.: Voter turnout				
Mean years of education	0.0111*	-0.0722***		
	(0.006)	(0.022)		
Upper secondary education			0.0170	-0.533***
			(0.040)	(0.193)
Municipality FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes
Province-Year FE	Yes	Yes	Yes	Yes
Observations	11722	11722	11722	11722
R^2	0.894	0.885	0.894	0.889
N. clusters	5861	5861	5861	5861
F-Test of Excl. IV		162.9		139.7

Note: The dependent variable is voter turnout. Explanatory variable in column (1) and (2) is mean years of education in the population aged 18+. Explanatory variable in columns (3) and (4) is the share of individuals aged 19+ with at least an upper secondary degree. Other controls include: share of females, concurrent elections and share of the population in working age. In columns (3) and (4) we control for the share of females aged 19+ and the share of the population aged 19-64. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.3: Alternative specifications

	IV	IV	IV
	(1)	(2)	(3)
dep. var.: Voter turnout			
Mean years of education	-0.0923*	-0.0738	-0.0416**
	(0.049)	(0.045)	(0.021)
Municipality FE	Yes	Yes	Yes
LLM - Time FE	No	No	Yes
Time FE	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Province-Year FE	Yes	Yes	No
Observations	11722	11722	11664
R^2	0.886	0.892	0.924
N. clusters	5861	5861	5832
F-Test of Excl. IV	38.92	35.31	135.6

Note: The dependent variable is voter turnout. In column (1), the share of working-age population is replaced with the population shares in 12 age bins, from 18-25 to 75 and over. In column (2), the share of working-age population is replaced with the population shares for all age years between 18 and 100. In column (3), province-by-time fixed effects are replaced with the finer local labor market by time fixed effects. Others controls are: share of females, concurrent elections, average age of the population 18+, and average age – by – time in columns (1) and (2) only. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.4: Robustness checks - alternative definition of the instrument

	IV	IV
	(1)	(2)
dep. var.: Voter turnout		
Mean years of education	-0.132*	-0.0949**
	(0.070)	(0.040)
Municipality FE	Yes	Yes
Time FE	Yes	Yes
Other Controls	Yes	Yes
Province-Year FE	Yes	Yes
Observations	11714	11720
R^2	0.868	0.880
N. clusters	5857	5860
F-Test of Excl. IV	10.35	39.59

Note: The dependent variable is voter turnout. Column (1) reports IV estimates when the population age structure used to defined the instrument is 10 years lagged ($k = 10$). Column (2) reports IV estimates when population age structure is 5 years lagged ($k = 5$). Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.5: Education, urbanization, and north-south dimension

	IV	IV
	(1)	(2)
dep. var.: Voter turnout		
Mean years of education	-0.0640***	-0.0311
	(0.022)	(0.021)
Mean years of education x rural	-0.0178***	
	(0.003)	
Mean years of education x south		-0.130**
		(0.061)
Municipality FE	Yes	Yes
Time FE	Yes	Yes
Other Controls	Yes	Yes
Province-Year FE	Yes	Yes
Observations	11722	11722
R^2	0.887	0.874
N. clusters	5861	5861
F-Test of Excl. IV Mean years of education	87.31	82.88
F-Test of Excl.IV Mean years of education x rural	5585.6	
F-Test of Excl.IV Mean years of education x south		21.77

Note: The dependent variable is voter turnout; *rural* is a dummy which takes 1 for rural municipalities; *south* is a dummy which takes 1 for municipalities located in South Italy. In column (1) controls include an interaction between share of population in working age and the dummy rural, while in column (2) an interaction between share of population in working age and the dummy South. Other controls are: share of females, concurrent elections and share of the population in working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.6: Education, voter turnout, social capital and income

	IV	IV	IV
	(1)	(2)	(3)
dep. var.: Voter turnout			
Mean years of education	-0.180***	-0.151***	-0.0543**
	(0.067)	(0.053)	(0.024)
Mean years of education x generalized trust	0.173**		
	(0.081)		
Mean years of education x donation		0.266**	
		(0.110)	
Mean years of education x GDPpc			-0.0326***
			(0.012)
Municipality FE	Yes	Yes	Yes
Time FE	Yes	No	No
Other Controls	Yes	Yes	Yes
Province-Year FE	Yes	Yes	Yes
Observations	11722	11572	11722
R^2	0.880	0.878	0.887
N. clusters	5861	5786	5861
F-Test of Excl. IV Mean years of education	89.10	99.23	81.65
F-Test of Excl.IV Mean years of education x generalized trust	95.55		
F-Test of Excl.IV Mean years of education x donation		104.6	
F-Test of Excl.IV Mean years of education x GDPc			745.9

Note: The dependent variable is voter turnout. Variables *generalized trust*, *donation* and *GDPpc* have been normalized in the interval [0,1] and interacted with Edu_{mt}. In column (1) controls include an interaction between the share of population in working age and the *generalized trust*; in column (2) an interaction between the share of population in working age and *donation*; in column (3) an interaction between the share of population in working age and *GDPpc*. Other controls are: share of females, concurrent elections and share of the population in working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.7: Education, voter turnout, charges of malfeasance and crime

	IV	IV
	(1)	(2)
dep. var.: Voter turnout		
Mean years of education	0.0242	0.0263
	(0.037)	(0.045)
Mean years of education x RAP	-0.223**	
	(0.111)	
Mean years of education x Crime		-0.279**
		(0.139)
Municipality FE	Yes	Yes
Time FE	Yes	Yes
Other Controls	Yes	Yes
Province-Year FE	Yes	Yes
Observations	11722	11722
R^2	0.872	0.873
N. clusters	5861	5861
F-Test of Excl. IV Mean years of education	93.84	109.4
F-Test of Excl.IV Mean years of education x RAP	63.53	
F-Test of Excl.IV Mean years of education x crime		90.50

Note: The dependent variable is voter turnout. Variables RAP (request to remove parliamentary immunity) and *crime* (degree of mafia infiltration) have been normalized in the interval [0,1] and interacted with Edu_{mt}. In column (1) controls include an interaction between the share of population in working age and RAP; in column (2) an interaction between the share of population in working age and *crime.* Other controls are: share of females, concurrent elections and share of the population in working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.8: Education, voter turnout, judicial inefficiency, corruption and infrastructure

	IV	IV	IV
	(1)	(2)	(3)
dep. var.: Voter turnout			
Mean years of education	0.0192	0.0589	-0.256***
	(0.035)	(0.042)	(0.079)
Mean years of education x judicial inefficiency	-0.292**		
	(0.124)		
Mean years of education x corruption index		-0.281**	
		(0.117)	
Mean years of education x infrastructure			0.372***
			(0.122)
Municipality FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Province-Year FE	Yes	Yes	Yes
Observations	11572	11722	11722
R^2	0.869	0.861	0.863
N. clusters	5786	5861	5861
F-Test of Excl. IV Mean years of education	82.66	108.7	102.0
F-Test of Excl.IV Mean years of education x judicial inefficiency	66.68		
F-Test of Excl.IV Mean years of education x corruption index		80.18	
F-Test of Excl.IV Mean years of education x infrastructure			108.0

Note: The dependent variable is voter turnout. Variables *judicial inefficiency*, *corruption index* and *infrastructure* (endowment of infrastructures) have been normalized in the interval [0,1] and interacted with Edu_{mt}. In column (1) controls include an interaction between the share of population in working age and *judicial inefficiency*; in column (2) an interaction between the share of population in working age and *corruption index*; in column (3) an interaction between the share of population in working age and *infrastructure.* Other controls are: share of females, concurrent elections and share of population in working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.9: Education, blank and invalid ballots

	Blank ballots	Invalid ballots
	(1)	(2)
Mean years of education	0.0331***	0.0344***
	(0.007)	(0.008)
Municipality FE	Yes	Yes
Time FE	Yes	Yes
Other Controls	Yes	Yes
Mean of outcome	0.0341	0.0659
Province-Year FE	Yes	Yes
Observations	11722	11722
R^2	0.900	0.914
N. clusters	5861	5861
F-Test of Excl. IV	162.9	162.9

Note: IV estimates. The dependent variable in column (1) is the share of voters casting a blank ballot. The dependent variable in column (2) is the share of voters casting a blank or an invalid ballot. Other controls are: share of females, concurrent elections and share of population in working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.10: Descriptive statistics - Itanes survey

	mean	sd	min	max	p50
Voter turnout	0.929	0.257	0	1	1
Years of education	9.949	3.838	0	17	8
Female	0.503	0.500	0	1	1
Year of birth	57.08	17.37	5	90	58
Year of birth (sq)	35.60	19.68	0.250	81	33.64
Observations	9322				

Note: Descriptive statistics on individual-level data collected by the Italian National Election Studies (Itanes) for the national elections of 2001, 2006, and 2008. Voter turnout is a dummy variable equal to 1 if the respondent declares he/she voted in national election and zero otherwise.

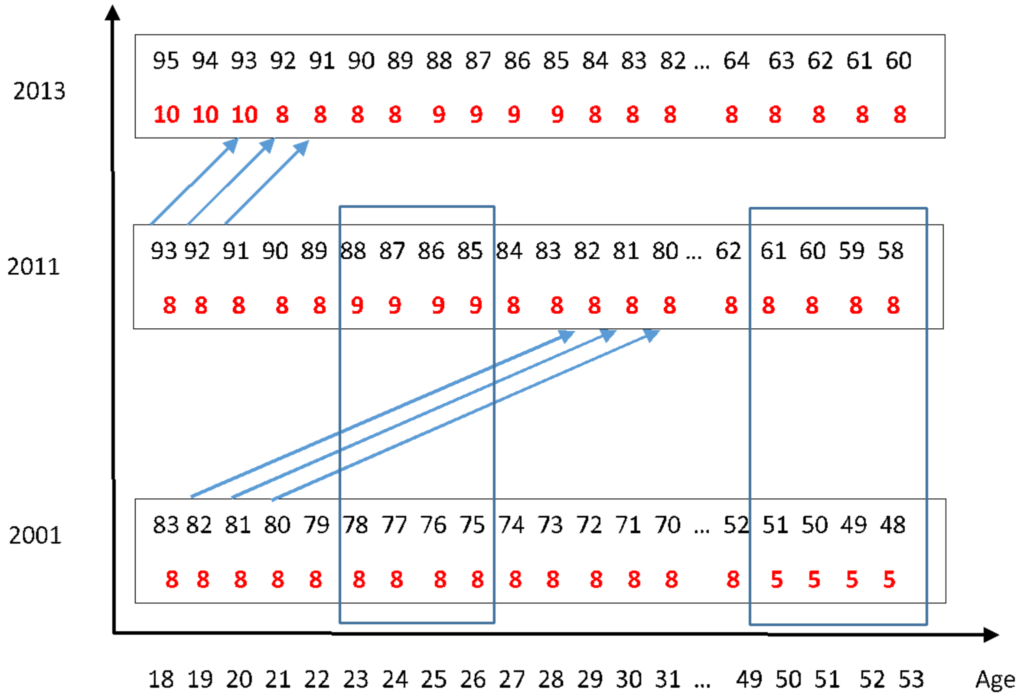
Table 1.11: Education and voter turnout - individual level analysis

	OLS	IV
	(1)	(2)
dep. var.: Voter turnout		
Years of education	0.00466***	-0.0498***
	(0.001)	(0.018)
Time FE	Yes	Yes
Region FE	Yes	Yes
Other Controls	Yes	Yes
Region-Year FE	Yes	Yes
Observations	9322	9322
N. clusters	85	85
F-Test of Excl. IV		12.93

Note: The dependent variable is voter turnout, a dummy variable equal to 1 if the respondent declares he/she voted in national election and zero otherwise. In column (2) *years of education* is instrumented by the compulsory number of years of education established by law for the cohort the individual belongs to. Other controls are: gender and a second order polynomial in the year of birth. Standard errors clustered by cohort. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

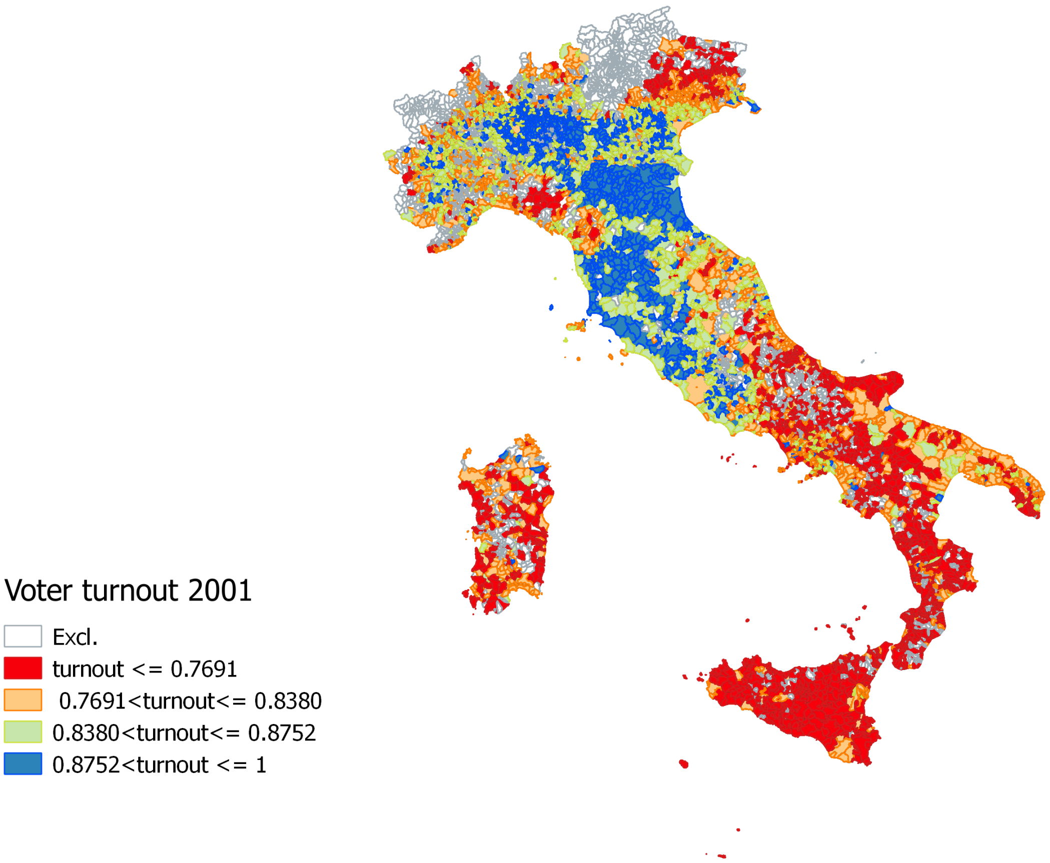
Figures

Figure 1.1: Reforms of compulsory education



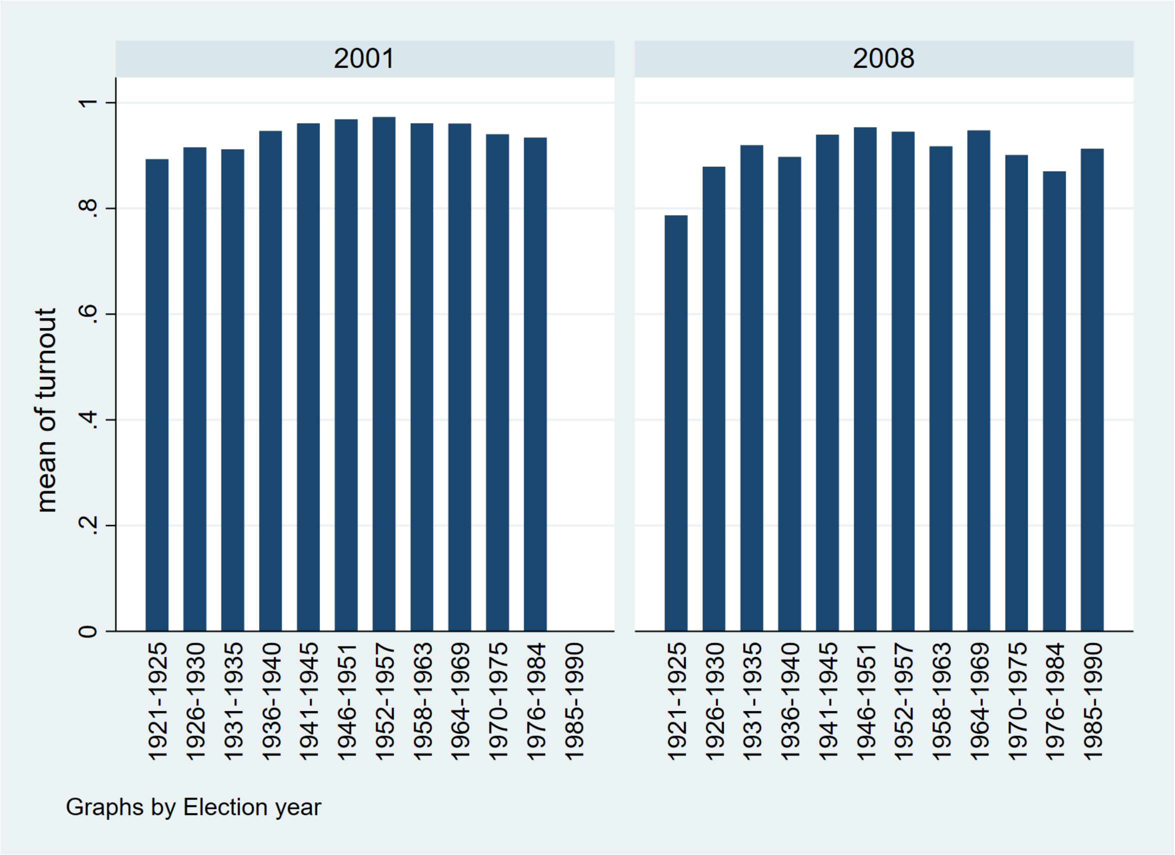
Note: in red is the length of compulsory education and in black the corresponding cohort of birth. The structure by cohort of the population aged 18+, who is eligible to vote for the Lower House, changes over time (2001, 2011, 2013). Arrows highlight how a given cohort "proceeds" in the structure by cohort of the population prevailing at each calendar year. Blue rectangles highlight the age-years for which there is a variation in the length of compulsory education between calendar years

Figure 1.2: Voter turnout across Italian municipalities



Note: voter turnout across Italian municipalities in parliamentary elections of 2001 (source: Ministry of Interior).

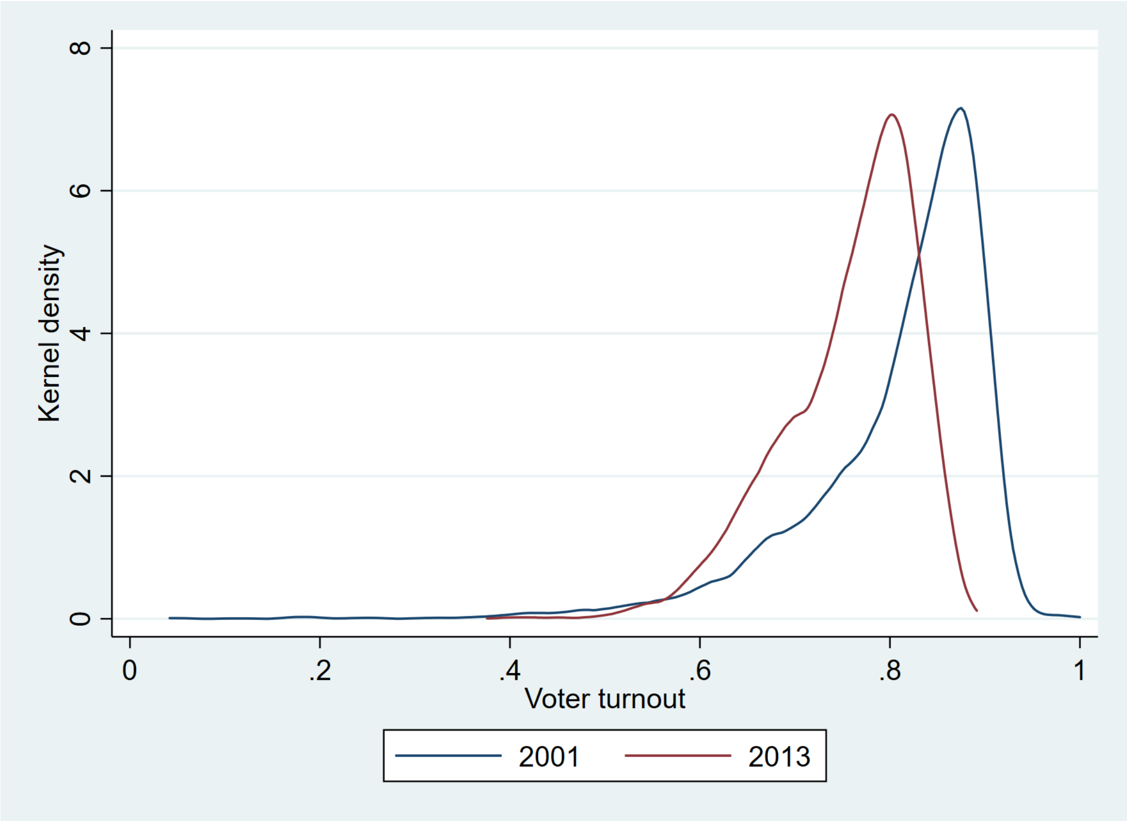
Figure 1.3: Voter turnout by cohort and election - Itanes survey



Note: Self-reported turnout rate by group of cohorts among the population aged 18+, in 2001 and 2008 (source: Itanes data).

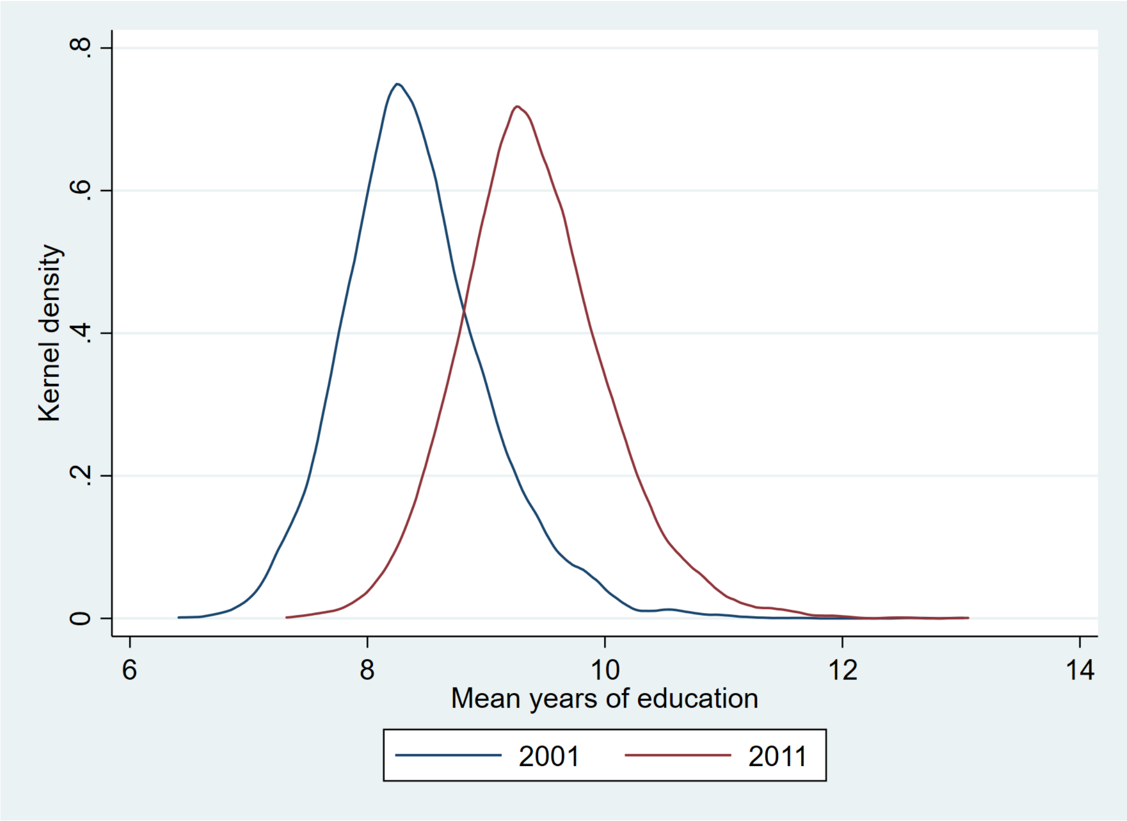
Appendix

Figure 1.4: Voter Turnout across municipalities. By year



Note: Distribution of turnout rate across municipalities, in 2001 and 2013 (source: Ministry of Interior - Archivio Storico delle elezioni).

Figure 1.5: Mean years of education across municipalities. By year



Note: Distribution of mean years of education across municipalities, in 2001 and 2011 (source: our elaboration on census data, ISTAT).

Table 1.12: Description of Variables

Variable	**Definition**	**Source**
Voter turnout	Share of eligible voters who cast a vote	Ministry of Interior, Archivio storico delle elezioni
Blank ballots	Share of voters who cast a blank ballot	Ministry of Interior, Archivio storico delle elezioni
Invalid ballot	Share of voters who cast a invalid ballot	Ministry of Interior, Archivio storico delle elezioni
Mean years of education	Mean Years of education of the population aged 18 and over	ISTAT
Upper secondary education	Index of upper secondary educational rate (19 and over)	ISTAT
Concurrent elections	Binary indicator (=1) if in the municipality were held concurrently local elections	Ministry of Interior, Archivio storico delle elezioni
Share of females	Share of females aged 18 and over	ISTAT
Working-age population (18-64)	Share of population in aged 18-64	ISTAT
Generalized Trust	Fraction of the individuals answering “Most people can be trusted” by region	EVS (2015)
Blood donations	Number of blood bags collected in the province per million inhabitants in 1995	Guiso et al. (2004)
GDP	GDP per capita at the province level in 2001	OECD Regional Economic dataset
South	Binary indicator (=1) if Southern municipality	ISTAT
Urban	Binary indicator (=1) if urban municipality	ISTAT
RAP	Fraction of deputies who have received at least a request to remove parliamentary immunity averaged at the electoral district	Golden (2007)

1.12 - Continued - Description of Variables

Variable	**Definition**	**Source**
Crime	Principal component of a factor analysis:mafia index, rates of crime against persons, properties, State, and other crimes per 100,000 inhabitants in year 2000 by province	Calderoni (2011), ISTAT
Judicial Inefficiency	Number of years to complete a trial	Guiso et al. (2004)
Corruption Index	Golden-Picci index	Golden and Picci (2005)
Infrastructure	Endowment of economic infrastructures in 1997, province level	Ecoter (2000)

Table 1.13: First Stage

	(1)	(2)
	Mean years of education	Upper secondary education
Mean years of compulsory schooling	1.239***	0.131***
	(0.097)	(0.011)
Municipality FE	Yes	Yes
Time FE	Yes	Yes
Other Controls	Yes	Yes
Province-Year FE	Yes	Yes
Observations	11722	11722
R^2	0.985	0.981
N. clusters	5861	5861
F-Test of Excl. IV	162.9	139.7

Note: Other controls are: share of females, concurrent elections and share of the population in the working age. Standard errors clustered at the municipality level. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 1.14: First stage - Individual level analysis

	OLS
	(1)
dep. var.: Years of education	
Years of compulsory schooling	0.259***
	(0.072)
Time FE	Yes
Region FE	Yes
Other Controls	Yes
Region-Year FE	No
Observations	9322
N. clusters	85
F-Test of Excl. IV	12.93

Note: The dependent variable is individual years of education. Other controls are: gender, and a second order polynomial in the year of birth. Standard errors clustered by cohort. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Chapter 2

Dialects, human capital and labour market outcomes

Elona Harka

2.1 Introduction

Italy is very rich in local dialects, which for centuries were the mother tongue of the majority of the population. It was only in the 1860s, with the political unification of the country, that the Italian language was imposed as the national language, although only a small fraction of the citizens of the newly established state spoke Standard Italian, origins of which can be traced back in the fourteenth century from the Florentine variety. What makes Italy stand out is that unlike other countries, dialects are not varieties of the Italian language, and they differ, considerably from Standard Italian. As Maiden and Parry (1997) suggest "there is probably no other area of Europe in which such a profusion of linguistic variation is concentrated into so small a geographical area". This makes Italy a unique case to study the importance of dialects, their distance from the national language on individual education and economic outcomes.

Although the role of dialects in individual success in education and labour market outcomes might matter, the literature on this topic is thin on the ground. This papers aims to fill this gap in the literature by investigating whether historical linguistic differences within Italy have had long-lasting effects on current outcomes. By adopting the view that Standard Italian was a "second" language and historical dialects the mother tongue of the majority of the population, the central hypothesis of this paper is that linguistic similarity to Standard Italian eases educational achievement and entry in occupations requiring high language skills. To empirically assess the consequences of linguistic distances, I exploit novel linguistic quantitative data from Hans Goebl (Goebl, 2008) derived from The Linguistic and Ethnographic Atlas of Italy and Southern Switzerland, which surveyed spoken dialects in some Italian municipalities at the beginning of the 20th century.

Dialects are not randomly distributed across space, and a variety of factors, from geography to historical institutions, may have shaped the extent of their similarity to Standard Italian. I suggest that, conditional on historical and geographical controls, dialects within province are as good as randomly assigned. The main empirical strategy employs a historical province fixed effects model and conditions to a battery of pre-determined controls. To test whether dialects, as spoken at the beginning of the 20th century, have consequences today, I use aggregate census data at the municipality level and construct a set of historical controls to account for possible factors affecting dialect similarity to the national language. For a subset of municipalities, I examine the effect of dialect similarity to Standard Italian on schooling, by also employing individual data from Labour Force Survey, conducted by the National Institute of Statistics. The central finding of this paper is that conditional on historical, geographical and contemporary controls, municipalities characterized from a historical dialect closer to the Standard Italian have better educational outcomes, and have been specializing in occupations requiring high skills in Standard Italian.

While local dialects were historically the primary language of the population, nowadays the situation is reversed and their use is limited. I suggest that the results of this paper are not driven only by dialect speakers, and propose two plausible mechanisms explaining the long-lasting effects of dialects: intergenerational transmission of language skills and intergenerational transmission of educational attainment. By building my hypothesis upon the literature on international migration (e.g., Chiswick and Miller, 2001) and literature assessing the consequences of language policy (e.g., Laitin and Ramachandran, 2016), I assume that the lower the distance from the national language, the lower the cost to acquire linguistic skills, and therefore the lower the costs to obtain education and entry in occupations intensive in language skills. Thus, the lower the linguistic distance, the lower were language deficiencies in Standard Italian and the better was education of older generations. Proficiency in Italian is transmitted from parents to children and educational levels persist across generations. Therefore, I conjecture that individuals living in communities characterized by a dialect that is closer to the Standard Italian will have better linguistic skills today. To assess whether dialects affect current linguistic skills, I exploit administrative data from the National Institute for Evaluation of Education (INVALSI) on standardized test scores of pupils enrolled in primary schooling. I find that children living in communities characterized by a dialect that is more similar to the Standard Italian, perform better in language test, regardless of whether they are dialect speakers or not. For dialect-speaking pupils, the effect is stronger, suggesting the linguistic similarity to Standard Italian might affect their academic performance directly. To test the central hypothesis that language affects academic performance in general and not only linguistic skills, I provide evidence that dialect similarity is positively associated with math scores.

This study adds to a recent literature that studies dialects and economic outcomes (e.g., Falck et al., 2012; Lameli et al., 2015; Falck et al., 2018; Yao and van

Ours, 2018; Grogger, 2011; Bian et al., 2019). Related literature, with exception of Yao et al. (2016), however, studies aspects like internal migration, earnings, employment probabilities, and do not focus on human capital. The only study that assesses the role of dialects on academic performance is Yao et al. (2016). However, Yao et al. (2016) focus only on schooling performance of young children, and put the emphasis mostly on dialects speaking and their linguistic distances are at a more aggregate level. This study indeed provides evidence from aggregate census data that linguistic similarity between dialects and the official language has had consequences on the educational attainment of whole communities. To a lesser extent this paper also contributes in the literature assessing the role of language policy on socioeconomic development (e.g.,Laitin and Ramachandran, 2016).

The remainder of this paper is organized as follows: Section 2.2 summarizes related literature. Section 2.3 provides some historical background on linguistic situation in Italy and describes the linguistic data. Section 2.4 introduces historical and contemporary data. In Section 2.5, I introduce the main empirical strategy and results. In Section 2.6 I propose possible mechanisms why historical dialects affect current outcomes. Section 2.7 presents the robustness checks. Section 2.8 concludes.

2.2 Related Literature

The economic effects of dialects have been largely ignored in the economics literature. However, recently, economics literature has investigated its economic effects. [1] Existing studies examine on how differences in dialects affect migration flows (Falck et al., 2012); bank performance (Bian et al., 2019); intra-national trade (Lameli et al., 2015); commuting flows (Persyn, 2017). While this literature fo-

[1] see (Suedekum, 2018) for a survey.

cuses on inter-dialectal similarity, another strand of the literature focuses on the effects of speaking dialects (e.g.,Yao and van Ours, 2018) and proficiency in the standard language (e.g.,Gao and Smyth, 2011).

Falck et al. (2012) employ linguistic data derived from a language survey carried out in 19th century and show that migration flows today are higher between areas with more similar dialects. Lameli et al. (2015) find that dialect similarity influence trade within Germany which they measure as the volume of shipments for each pair of region. Persyn (2017) suggest that commuting flows are lower between areas with a more distinct dialect.

Falck et al. (2018) focus on internal migrants in Germany and assess the wage premium individuals demand to overcome cultural dissimilarity measured as dialect distance between German counties. Bian et al. (2019), by combining data at the county level of Chinese dialects with data on commercial banks, suggest that dialect similarity between the chairman and the CEO is positively associated with proxies of bank performance like return on assets and equity. Grogger (2011) and Grogger (2018) exploiting US speech data investigate the role of within language variation on wages. Grogger (2011) study finds that black workers speaking a racially distinctive dialect earn less than whites with similar skills. Grogger (2018) findings suggest that both racial and regional speech patterns are associated with wages. Gao and Smyth (2011) investigate the effect on earnings of speaking Standard Mandarin for rural-urban migrants in China. They employ survey data on respondent self-reported proficiency in Standard Mandarin and suggest that speaking Standard Mandarin has a positive impact on earnings. Authors find also gender differences on economic return of language fluency, which has a large positive effect on female earnings but it is statistically insignificant for males. Yao and van Ours (2018) and Yao and van Ours (2019) explore whether there is association between daily dialect speaking and hourly wages in the Netherlands. Yao and van Ours (2018) suggest daily dialect speaking is negatively correlated with wages for men, but they do not find a statistically significant result for females. In a more recent study Yao and van Ours (2019) by instrumenting dialect speech with the geographical distance from the municipality of Haarlem, where the purest form of Standard Dutch is spoken, confirm their previous findings. Dovì (2019) uses surveys containing recorded data on proficiency levels of the respondents to assess the role of Mandarin proficiency on employment probabilities in China. By exploiting as instruments for language proficiency, whether Mandarin is the primary language spoken outside the workplace and its status of lingua franca of China, the author finds that fluency in the official language is positively associated with employment probabilities.

Dialects have also been associated with educational outcomes. Yao et al. (2016) investigate the role of dialect speaking on educational achievement on young children in the Netherlands. Their findings suggest that speaking dialect at home adversely affect language test scores of boys. The negative effect is small but increases as the linguistic distance between dialect and Standard Dutch increases.

Language proficiency has also been related with other aspects of economic behaviour. For instance, Wang et al. (2016) finds that proficiency in Mandarin is positively associated with consumption expenditures.

This study also is related to the literature on the role of the language policy. Laitin and Ramachandran (2016) findings suggest that in countries that kept the language of the colonizer after independence distance and exposure to the official language influence socioeconomic development.

2.3 Background and Linguistic data

2.3.1 Background

There are 35 living languages in Italy, according to the Ethnologue. Among them, 28 are indigenous languages. (Eberhard et al., 2019). Italian is the formal language of Italy and derives from the Tuscan dialect, in particular, the Florentine variety. Although it originated in the fourteenth century, Italian was a literary language, and its consolidation as a spoken language happened only in the 20th century (Wieling et al., 2014). In fact, for centuries the population used their local dialects. “Not only was Italian used far less than these dialects, it was also less important than Latin at Rome, French at Turin, and Spanish in Naples, Sicily, and Sardinia.”(Smith, 1988, p.2). Dialects were not only used by lower classes. They were even employed in institutional occasions. For instance, in the decades prior to the unification of Italy, in Venice dialect was used in courts and by politicians (De Mauro, 1991, p.32). A necessary clarification of the term dialect in this context. The term here does not refer to varieties of the Italian language. There are substantial differences between local dialects and Standard Italian. “Italian dialects differ from literary Italian and among themselves so much that one dialect may be unintelligible to the speaker of another dialect” (Lepschy and Lepschy, 1988, p.13).

If we turn to quantitative measures and consider, for example, linguistic data on the municipalities analyzed in this paper, the linguistic distance among some dialects might be higher than the distance between Italian and French. Table 2.1 displays linguistic similarity between pair of municipalities. Row one shows the linguistic similarity between a municipality in Sardinia region and one in Aosta Valley. Linguistic similarity between them is lower than that between Standard French and Standard Italian. The linguistic similarity between two capital provinces as Palermo and Bologna is almost the same as that between French and

Italian.[2]

After the political unification of Italy in the 1860s, writer Alessandro Manzoni in his report of 1868 "On the Unity of Language and the Means to Propagate it" [3] to the Minister of education proposed that the national language to be adopted was contemporary Florentine variety. However, at the political unification, around 2.5 percent of the population spoke Italian (De Mauro, 1991). Despite the majority of the population had no contact with Italian, the educational system engaged in diffusing literacy skills in Tuscan variety and eradication of Italo-Romance varieties (Guerini, 2011). "The fight against dialects was considered central to the educational process, because schoolchildren (and illiterate adults) in learning Italian had to suppress the interference of the dialect, which was their native tongue"(Lepschy and Lepschy, 1988, p.28).

Promotion of the national language also continued during the 20th century, by reaching more severe dimensions during the Fascist regime. Although the educational reform of the 1923 recognized the importance of dialects as a starting point in the educational system (Lepschy and Lepschy, 1988) one of the objectives of the fascist policies was the repression of dialects (Tosi, 2001). However, according to Tosi (2001), dialects were still predominant, and if a language shift from the dialect to the national language occurred during that period, it was restricted to the upper and middle class in the main urban centers.

Dialects did not disappear and according to De Mauro (1991) in 1951, 13 % of the individuals were using only dialects and 18.5 % only Italian. Furthermore, around 63.5 % normally used their dialects in most circumstances.

Differently from the past, dialects are not so widely used in Italian society today. In fact according to the inquiries on the use of the Italian language, dialects, and foreign language conducted by the National Institute of Statistics in 2015 around

[2] Figure 2.8 in the appendix shows the geographical location of these municipalities.

[3] In Italian Dell'unita della lingua e dei mezzi di diffonderla

45.9% declares that at home speaks mainly Italian, and around 32.2% reports to be using both. However, there are regional peculiarities. For instance, in regions like Veneto around 30.6 % reports to speak mainly dialect at home.

2.3.2 Linguistic Data

To empirically assess the role of historical language barriers, I exploit linguistic research on dialects surveyed in some Italian municipalities at the beginning of the 20th century. I rely on a measure of linguistic similarity from Standard Italian elaborated from Hans Goebl (Goebl, 2008). The work of Goebl is based on The Linguistic and Ethnographic Atlas of Italy and Southern Switzerland (AIS, Sprach-und Sachatlas Italiens und der Südschweiz) by Karl Jaberg and Jacob Jub. The Atlas was published between 1928-1940 and contains 1705 linguistic maps. Surveys on the field were concluded in 1928 (Goebl, 2016). Figure 2.1 displays some of the localities surveyed in Lombardy region in North Italy. Figure 2.2 shows how the word "the boys" ("i ragazzi" in Italian) is pronounced in each locality. In this example, the lexical choice shows that even in small geographic areas, there are distinctions between dialects.

I reconstructed municipalities surveyed in AIS according to administrative boundaries of 2011. Overall the sample consists of 338 municipalities. My measure of linguistic similarity to Standard Italian refers to the index of relative identity value (indice relativo d'identità in Italian) (Goebl, 2008). In particular it is calculated from the number of pairwise matchings and mis-matchings of linguistic elements called *taxates* (Goebl, 2018). Formally the index is presented as below :

$$\mathrm{RIV_{jk}} = 100\frac{\sum \mathrm{COI}(i)_{\mathrm{jk}}}{\sum \mathrm{COI}(i)_{\mathrm{jk}} + \sum \mathrm{COD}(i)_{\mathrm{jk}}} \tag{2.1}$$

where $\mathrm{RIV_{jk}}$ is the Relative Identity Value and takes values $0 \leq \mathrm{RIV_{jk}} \leq 100$. j is the reference site and k the site to be compared with j while i represents one of the p working maps. COD$(i)_{\mathrm{jk}}$ and COI$(i)_{\mathrm{jk}}$ are respectively co-difference and co-identity between two *taxates* (Goebl, 2018).

Figure 2.3 shows the geographic location of the municipalities of the sample. Figure 2.4 through Voronoi polygons displays the spatial patterns of linguistic similarity to Standard Italian. Tuscan dialects display the highest degree of similarity. Table 2.2 reports some descriptive statistics of the index of relative identity value. Figure 2.9 in the appendix shows its distribution.

2.4 The data

This section introduces the main variables of the analysis at the municipality level.

2.4.1 Contemporary variables

The main outcome of interest for this study is the educational attainment at the municipality level which is measured as the mean years of education in the municipality in 2011, or as upper secondary educational attainment rate. I construct mean years of education in the municipality from grouped administrative data of the National Census on the highest educational attainment level of the resident population aged 18 and over. Upper secondary educational attainment rate is defined as the percentage of individuals aged 19 and over who have completed at least upper secondary education. The source for these data is National Institute of Statistics (ISTAT). Concerning the labour market, the outcome of interest is the percentage rate of high skilled workers. This category comprises legislators, entrepreneurs and top management, intellectual, scientific and highly skilled professions, and technical professions. These data are available on the *ISTAT's 8000 Census* portal, which contains indicators for all Italian municipalities.

Demographic data at the municipality level like ageing index, gender ratio, demographic density, the share of the resident population in non-urban areas in each municipality are also drawn from *8000 Census* portal. Mean age of the population is computed from detailed grouped census data of population by age. Geographic data like latitude, longitude, altitude area, and other municipality characteristics are drawn from ISTAT. To take into account that the Italian language derives from the Tuscan dialect, I control for the geographic distance from Florence. Table 2.19 provides a detailed description of these variables.

2.4.2 Historical Variables

A variety of historical factors may have affected linguistic similarity between the dialect prevalent in an area and Standard Italian and simultaneously drive educational and labour market outcomes nowadays. To account for possible confounders which drove dialect formation and contemporary economic outcomes, I add a battery of pre-determined controls. [4]
First, I control for literacy rates, [5] because as discussed in Section 2.3 educational system engaged in diffusing literacy skills in the national language. Data on literacy rates are drawn from the Population Census of 1911. Second, I control for the presence of libraries in the municipality, because access to books written in Standard Italian might have altered the degree of similarity to the Standard language, but also fostered educational achievement in an area. The source for these data is a publication of 1893 on the statistics of the libraries. Third, I control for the presence of transportation networks like railways and ports and communication services like the telegraph and postal offices. Railway networks or ports might

[4] This set of controls is constructed according to the historical boundaries of the municipalities.

[5] School supply could be a possible key variable to be included in the analysis. Historical data at the municipality level on primary schooling are available from a report by the Italian Statistical Office for the academic year 1862-1863. The report contains information on the numbers of schools and teachers, the number of pupils enrolled and expenditures. Based on these data, it is possible to construct pre-determined measures of school supply; however, for around 25 percent of municipalities in the sample, this information is missing, mostly because they joined Italy in a later period. As a robustness check, I estimate regressions controlling for the number of schools at the municipality level and student/teacher ratio. For the municipalities with missing information, I substitute these variables with zero and code a dummy, which I include in all estimations. Accounting for school supply does not affect the baseline results.

have fostered economic exchange between municipalities that could have possibly affected not only local development, but also might have affected dialect similarity between areas, and therefore altered the similarity with Standard Italian. These indicators are drawn from the dictionary of municipalities of the Kingdom of Italy published in 1874 by the Italian Ministry of Interior. Table 2.19 provides a detailed description of these variables, while Table 2.3 presents some descriptive statistics. [6]

Finally, I take into account former administrative boundaries. Falck et al. (2012) in the context of Germany, argue that former independent territories have contributed to linguistic evolution. Similarly, Italy for centuries have been divided into small independent States; hence, I also control for historical state affiliation. [7]

Besides digitizing municipality level data, for an alternative model specification, I construct historical variables at the province level. I digitized data from the Statistical Yearbook of Italian provinces of 1872 (from now on referred as the Yearbook), National census of 1871 and a publication titled "L'Italia Economica nel 1873".[8]

To control for historical conditions in education, I include in the analysis, the following measures for primary schooling: enrollment rates, student-teacher ratio, expenditures per student, and the number of primary schools. I control also for the historical rate of the population employed in high skilled occupations. Here, I

[6] There is no information on these data for around 6 percent of the municipalities of the sample because they joined Italy after the First World War. For these municipalities, I assign zero to these variables and code a dummy equal to one if the information is missing and zero otherwise, which I include in all estimations.

[7] I control for historical state affiliation in 1850. Municipalities of the sample in 1850 were part of 8 different historical states namely the Kingdom of Sardinia, Kingdom of Two Sicilies, Austria, Kingdom of Lombardy-Venetia, Papal State, Grand Duchy of Tuscany, Duchy of Modena (and Massa Carrara), Duchy of Parma, Piacenza, and Guastalla.

[8] For the province of Rome some information is missing in the Yearbook since it joined Italy in 1870. The information is missing also for the sample of municipalities joining Italy after the First World War. Accordingly, I code a dummy also for these variables which I include in all estimations.

define high skilled workers those people covering the following occupations: law, science and humanities, medicine and education.

Journals published in the Italian language could be a possible factor affecting dialect similarity to Standard Italian. The Yearbook contains information about the absolute number of journals for each province. Almost all of them were published in Italian (Yearbook).

To construct proxies for income, I use data on tax collection and eligible voters. The Yearbook contains information on a variety of taxes collected in each province in 1870. Among them, I use the share of direct taxes collected per inhabitant. Eligible voters per 100 inhabitants in political elections of 1870 are used also as a proxy for income since the electoral law in force in the period limited the right to vote based on requirements on minimum age, literacy and wealth. Finally, I construct proxies for local development. As a proxy for financial development, I construct a dummy indicating the presence of financial institutions and banks. As a proxy for financial instruments, I include the number of postal money orders per 1000 people and their value in thousands *"lire"* per inhabitant [9]. Additional proxies for development include the density of post offices, national and provincial roads per square km and province population density. Table 2.4 shows some summary statistics of this set of controls, while Table 2.21 in the Appendix provides details on the sources of the raw data and how variables are constructed.

2.5 Empirical Strategy

To assess whether dialects, as surveyed at the beginning of the 20th century, have long-lasting effects on educational attainment and labour market, I estimate the

[9] I normalize these variables by the province population in 1871 ,which is drawn from National census of Italy

following reduced form regression model:

$$Out_{mp} = \alpha_0 + \alpha_1 DIALECT_{mp} + X_{mp}\beta + \delta_p + \varepsilon_{mp} \tag{2.2}$$

Where Out_{mp} is the outcome of interest at municipality m in province p. The main explanatory variable $DIALECT_{mp}$ is dialect similarity to Standard Italian of municipality m in province p measured by the index of relative identity value. X_{mp} is a vector of control variables including, geographical, demographic, and historical variables, historical state affiliation, and other municipality characteristics.

To account for correlation between dialect similarity and other local characteristics I exploit within province variation. δ_p denotes historical province fixed effects. I avoid using today's administrative borders because historical borders could have possibly affected dialects and current province borders could be an outcome of linguistic similarities. In particular, I use the province administrative borders of 1871. ε_{mp} denotes the standard errors. [10]

The coefficient of interest is α_1. I expect to be positive and statistically significant, showing that conditional on preexisting conditions and the set of other controls - linguistic proximity from Standard Italian increases educational attainment and access in occupations requiring high skills in Italian. The main assumption of the empirical strategy is that conditional on preexisting conditions historical dialects within province are as good as randomly assigned.

In an alternative specification, I substitute province fixed effects with historical control variables at the province level and estimate the following model:

$$Out_{mp} = \alpha_0 + \alpha_1 DIALECT_{mp} + X_{mp}\beta + W_p\theta + \varepsilon_{mp} \tag{2.3}$$

[10] As a robustness check, I use Conley (1999) adjusted standard errors to deal with spatial correlation. Results (see Table 2.22) remain significant under different distance cutoff.

Where W_p is a vector of controls at the province level. W_p includes historical socioeconomic variables like education, financial and economic development, and proxies for income.

2.5.1 Results

This section presents the results on the link between linguistic proximity to Standard Italian and contemporary outcomes. First, I report the baseline results of this paper based on aggregate municipality level data. For a sub-set of municipalities, I provide some further evidence with individual-level data on the influence of dialects on schooling.

Baseline Results

Table 2.5 displays the main results on the effect of dialect similarity on mean years of education under different model specifications. In all models, I include historical state fixed effects. Under different specifications, I find that dialect similarity is positively associated with mean years of education and statistically significant at 5 percent level of significance. Column (2) shows the results when I exploit within historical state variation and control for contemporary and geographical variables. Column (3) reports the results of the model without province fixed effects but including historical, geographical, and contemporary controls at the municipality level as well as historical controls at the province level. Column (4) documents estimations of the preferred model, which exploits within province variation. According to the model in column (4) ten percentage point increase in dialect similarity is associated with around 0.26 percentage point increase in mean years of education. Put it differently; one standard deviation increase in dialect similarity accounts for 0.27 standard deviation increase in mean years of education.

Table 2.6 reports estimates when the dependent variable is upper secondary educational attainment rate. Column (3) shows the results of the alternative specification (equation 2.3). According to this model specification, a ten percentage point increase in linguistic similarity is associated with a 2 percentage point increase in the upper secondary educational attainment rate. The model with province fixed effects presented in column (4) confirms the positive relationship. A ten percentage point increase in the index of dialect similarity is associated with a 2.6 percentage point increase in the upper secondary educational attainment rate. Overall the reduced forms estimates under different model specifications indicate that historical dialects are associated with contemporary educational outcomes.

In Table 2.7 are shown estimations when the outcome of interest is the incidence of the employed in high skilled occupations. The central hypothesis to be explored here is that communities with a greater extent of linguistic similarity to Standard Italian have specialized in occupations requiring high language skills. OLS estimations presented in Table 2.7 suggest that linguistic similarity to the national language is positively associated with the rate of the employed in high skilled occupations. Column (3) reports the results of the alternative model with historical controls at the province level, while column (4) displays the estimations of the baseline model with historical province fixed effect. Under different specifications, I find a positive association between dialect similarity and the incidence of high skilled workers. In the preferred model with province fixed effects and historical controls at the municipality level, I document that ten percentage points increase in the index of relative identity value entails an increase of about 2.6 percentage point in the dependent variable.

A possible caveat of my findings is that I consider the relationship in only one fixed point of time. I conduct some further analysis, by considering the association in different periods. The classifications of occupations based on the level of skills and according to the administrative borders of 2011 it is available also for the national censuses of 1991 and 2001. In addition, this further analysis it is informative about the magnitude of the effect in different periods. Column (5) reports the OLS estimations of the baseline model with province fixed effect confirming the positive link between dialect similarity and labour market outcomes also in 2001. Column (6) displays the results for 1991. I find that a ten percentage point increase in the index of relative identity entails an increase of 3.1 percentage points in the incidence of high skilled workers. These results lend support to the hypothesis that a greater extent of dialect similarity to the national language eases entry in occupations intensive in language skills.

Additional findings

In this section, I present some further analysis by first modifying the index of the relative identity value by expressing it as quartiles of its distribution, and as a second step, I explore further the hypothesis that historical dialects affect educational achievement with individual-level data.

Table 2.8 displays the coefficients for each quartile dummy (the fourth quartile is the excluded category). The pattern of the coefficients suggests that years of education and upper secondary educational attainment rate decreases as linguistic similarity to Standard Italian decreases. The same pattern is observed when the dependent variable is the incidence of employed in high skilled professions.

To provide evidence on the link between dialects and educational outcomes with individual-level data, I exploit Labour Force Surveys (LFS) carried out from 2006 to 2011 by ISTAT. The surveys contain information on socio-demographic variables and labour market indicators. Overall, the sample consists of 231544 Italian citizens born between 1952 and 1996. Since the LFS survey is conducted in only 126 municipalities surveyed in AIS, it is not feasible to estimate the preferred model in equation 2.2 with province fixed effects. Analysis with aggregate data for educational outcomes suggests that model with province fixed effects and historical controls at the province level yields similar results. Therefore, I estimate the model in equation 2.3 with historical controls at the province level by substituting demographic variables with individual controls. Formally, I estimate the following model:

$$edu_{impt} = \alpha_0 + \alpha_1 DIALECT_{mp} + X_i\gamma + X_{mp}\beta + W_p\theta + \delta_t + \varepsilon_{impt} \tag{2.4}$$

where edu_{impt} is educational outcome for individual i at municipality m in province p at time t. X_i is a vector of individual-level controls including age, a second-order polynomial in age, and gender. δ_t are year fixed effects. The other controls are the same as in specification presented in equation 2.3.

I investigate the role of linguistic proximity on three educational outcomes - years of education, upper secondary educational attainment, and school dropout. Upper secondary education attainment is a binary indicator equal to one if the respondent reports that he/she has obtained at least a secondary education diploma. Dropouts are defined those individuals who did not comply with legislation on compulsory schooling. In particular, to define school dropouts, I refer to Law 1852/1962, which states that education was mandatory until students attain the middle school diploma. The first cohort affected from the Law was that of 1952. Here dropout is a binary indicator equal to 1 for individuals who have completed only elementary schooling and zero otherwise. For binary outcomes, I estimate a linear probability model.

I expect to find a positive and statistically significant effect of dialect similarity - indicating that conditional on preexisting conditions - living in a municipality displaying a greater extent of linguistic similarity to Standard Italian increases individuals' years of education and probability to obtain at least an upper secondary diploma. On the other hand, I expect to find a negative association on the probability of dropout from mandatory schooling.

Table 2.9 displays the results of this further investigation. Column (1) reports that there is a positive association between dialects and years of education. In column (2), I document that linguistic proximity to Standard Italian increases the probability to complete upper secondary education. A ten percentage point increase in dialect similarity entails an increase of around 5 percentage points in the probability to complete at least upper secondary schooling. As conjectured, living in a community displaying higher linguistic proximity to Standard Italian decreases the probability to drop out from mandatory schooling. According to the model shown in column (3), ten percentage points increase in dialect similarity decreases the probability of dropout from mandatory schooling by around 1,5 percentage points.

2.6 Possible mechanisms

As described in Section 2.3, dialects are not widely used today, and therefore they do not constitute barriers to communication. Moreover, one may argue that since Italian is widespread, they should not influence contemporary outcomes. This section discusses some possible mechanisms on how dialects may affect contemporary outcomes. Based on the linguistic background provided in Section 2.3, I assume that historically, the local dialects were the "primary" language of each community, while Standard Italian was the "second" language.

In the context of immigration studies, Chiswick and Miller (2001) in their framework of language attainment argue that efficiency in learning a second language depends on the linguistic difference between individuals' origin language and destination language. Chiswick and Miller (2001) and most recently Eduard Isphording and Otten (2013) suggest that the higher the distance between the language of the destination and origin, the higher the cost in the acquisition of the foreign language.

Laitin and Ramachandran (2016) in assessing the role of the language policy in countries retaining the language of the colonizer suggest that a greater distance from the official language decreases human capital due to increased costs in learning the language.

Similarly, in the present context, I assume that municipalities speaking a dialect with higher linguistic similarity to Standard Italian had to bear a lower cost in acquiring proficiency in the national language. In other words, the assumption states that individuals speaking, for instance, a Tuscan dialect had less difficulty to learn Standard Italian than individuals speaking, for instance, a Lombard dialect. The greater the extent of historical similarity, the lower the cost to acquire human capital and the lower the barriers to access to occupations intensive in Standard Italian.

To test the mechanism that linguistic distance influences the proficiency in Italian, I use data from the National Institute for Evaluation of Education (INVALSI) on academic achievement of pupils enrolled in the fifth grade of primary schooling in the academic year 2012-2013. The data are suitable to test the mechanism because they contain information on student citizenship, whether they speak dialects at home, and parents country of birth. INVALSI data are administrative data; thus representative of all Italian students enrolled in the fifth grade of primary schooling. In the sample, I include children with Italian citizenship whose both parents were born in Italy. In addition, I restrict it to those pupils reporting to speak at home Italian or a local dialect [11]; the final sample consists of 59307 observations. The outcome of interest is the score in the Italian test. Based on the above assumption, I hypothesize that children living in municipalities with a greater extent of linguistic proximity to the national language perform better in

[11] In INVALSI questionnaires, pupils indicate which language they speak most of the time at home. Each pupil was asked to report one of the following categories: Italian, a dialect, or a foreign language. I exclude from the sample pupils reporting to speak a foreign language. It has to be noted that from this question cannot be inferred whether a pupil speaks only dialect at home.

the Italian test.

The underlying mechanism, I propose is that of intergenerational transmission of linguistic skills. In the context of immigration studies, Casey and Dustmann (2008) suggest there is an association between parental fluency in the language of the destination country and proficiency of their children. I propose a similar mechanism in this setting to explain why dialects may have long-lasting effects even though their use is limited today. The higher the linguistic distance, the greater were the language deficiencies in Standard Italian of older generations. Linguistic skills are transmitted within families from parents to children. Therefore, children born from parents with low linguistic skills will have lower proficiency in Standard Italian, regardless whether they are dialects speaker at home or not. For dialect speakers pupils, the role of dialects might be more straightforward. Dialects are transmitted within the family from parents to children; therefore, dialect similarity to Standard Italian might directly affect their linguistic skills. Thus, I expect that the extent of dialect similarity to Standard Italian to have a stronger effect on children speaking dialects at home.

To test the above conjectures empirically, I estimate the preferred model with province fixed effects by substituting demographic variables at the municipality level with students characteristics. Formally, I estimate the following model:

$$Score_{imp} = \alpha_0 + \alpha_1 DIALECT_{mp} + X_i\gamma + X_{mp}\beta + \delta_p + \varepsilon_{imp} \qquad (2.5)$$

Where $Score_{imp}$ is the test score for student i enrolled in school in municipality m in province p. X_i is a vector of individuals controls, including gender, year of birth, and a second-order polynomial in the year of birth. All the other variable are the same as in the baseline specification. As a second step, I extend the model by including a binary indicator equal to 1 if student reports to speak dialect at home, and an interaction term between this binary indicator and linguistic similarity to Standard Italian.

Table 2.10 reports the results of both specifications. In column (1) I document that linguistic similarity to Standard Italian is positively associated with the Italian test score. A ten percentage point increase in the index of relative identity value is associated with an increase of 2.33 percentage point in the Italian test score. The positive effect is higher for students speaking dialect at home. Column (2) shows the estimates of the extended model. This further analysis suggests that for students speaking Italian at home, the effect is smaller and marginally significant. Indeed for children speaking dialect at home, the positive effect of linguistic similarity is stronger. A ten percentage point increase in the relative identity value is associated with an increase of around 4.48 percentage points in the Italian test score.

Since the central hypothesis of the paper is that dialects affect educational achievement in general, I estimate whether dialect similarity affects math scores. Though it might be counter-intuitive, as one may argue that mathematics involves just calculations, there is evidence that language could impact math scores. For instance, Abedi and Lord (2001) conduct two field studies to assess the role of students' linguistic background on their scores in math word problems. According to the authors, English language learners students perform worse than their counterpart proficient in English and moreover, the former benefited more from the simplification of the language of the test.

Therefore, I assess whether dialects affect math scores. Column (3) and (4) of Table 2.10 reports the estimation of both models when the dependent variable is the score in the math test. Surprisingly, the effect of linguistic similarity to Standard Italian is higher on math scores. According to the model in column (3) an increase of ten units in the index of relative identity value will cause an increase in the math test scores by 4 percentage points. For pupils speaking dialect at home according to the model in column (4), ten units increase in dialect similarity is associated with an increase of 6.87 percentage points in the math test score.

A second plausible mechanism is that of intergenerational transmission of educational achievement. Italy, has high levels of intergenerational persistence in schooling attainment (Checchi et al., 2013). Dialects might have affected schooling opportunities of older generations. Educational achievement is transmitted between generations, and therefore, children born from educated fathers attain more education.

2.7 Robustness checks

In this section, I assess the robustness of findings by: i) conducting placebo regressions and estimating the same models of the above sections with individual data for samples of immigrants, ii) producing a placebo test showing that for outcomes of interest in this study is the distance from Standard Italian that matters. iii) addressing possible issues of reverse causality iv) assessing robustness to omitted variable bias, v) ensuring that the results are not driven by any province in particular, vi) randomized inference.

Placebo regressions: The results of the previous sections point out that municipalities characterized by a historical dialect more similar to Standard Italian have better contemporary educational outcomes. If results arise due to spurious correlation, the positive and statistically significant relationship of linguistic similarity on educational attainment should arise for immigrants too. Historical dialects should not affect the educational outcomes of immigrants since Italy became a country of destination only a few decades ago when Italian was widespread. Therefore, I conduct placebo regressions with individual-level data from which it is possible to distinguish between Italian citizens and foreign nationals.

The first set of placebo regressions is carried out with Labour Force Survey data by estimating the model in (equation 2.4) for a sample of immigrants. The sample consists of 23376 individuals reporting to be foreign nationals. The older cohort in the sample is that of 1952. The choice of 1952 is made to keep the samples as comparable as possible, considering that immigration is a recent phenomena and immigrants are usually young. Table 2.11 displays the placebo regressions. Coefficients on dialect similarity are not statistically significant for all outcomes of interest.

It is possible to conduct the same placebo exercise also with INVALSI data since it contains information on pupils citizenship. The sample consists of 9823 pupils who are first or second-generation immigrants. Table 2.12 displays the estimations of the model in equation 2.5 for the sample of non-natives pupils. Column(1) reports model estimation when the dependent variable is the Italian test score. Column (2) displays the results for math score tests. For both outcomes coefficient on dialect similarity, it is not statistically significant.

Placebo test: To validate my hypothesis that for educational attainment and entry in high skilled occupations is the similarity from the official language that matters, I carry out a set of regressions by using as explanatory variable linguistic similarity to dialects spoken in big cities (i.e., large, rich, educated cities and big regional capitals). I estimate the preferred model with province fixed effects (equation 2.2) by employing as reference sites cities of Turin, Milan, Venice, Bologna and Naples (i.e., linguistic similarity to the dialects spoken in these cities). Results for the three outcomes of interest are shown respectively in Table 2.13, Table 2.14 and Table 2.15. For the three outcomes of interests, the coefficient on dialect similarity is not statistically significant for any of the above reference points, lending support to the central hypothesis of this paper that is the linguistic similarity to Standard Italian that eases educational achievement and entry in occupations requiring high language skills.

Reverse causality: Another possible concern is that the Tuscan dialect could have been chosen as Standard Italian due to the higher human capital in the area.

Under the assumption that human capital persists, this could raise concerns on reverse causality. To address this issue, I estimate the model with province fixed effect by using as a dependent variable a pre-determined measure of education, i.e., the literacy rate in 1911 and by controlling for geographical and historical variables. Results (Table 2.16) show that conditional on historical and geographical controls there is no statistically significant association between dialect similarity to Standard Italian and literacy rates in 1911.

Robustness to omitted variable bias: Employing historical dialects as the main explanatory variable and contemporary outcomes as the dependent variable has the advantage to limit issues regarding reverse causality. However, concerns about omitted variable bias arise despite, I include historical province fixed effects and control for historical conditions possibly affecting linguistic similarity to Standard Italian. To assess the robustness to omitted variable bias, I use the test suggested by Oster (2017). I follow the approach to calculate the bias adjustment effect β under the assumption that $\delta = 1$, that is equal selection and $R_{\text{max}} = 1.3\tilde{R}$. Table 2.17 report β for the regressions with historical province fixed effects. For the three outcomes of interest, the adjusted treatment effect has the same sign as the coefficients.

Sub-samples: To ensure that the results are not driven by any province in particular, I estimate the preferred model with province fixed effects by excluding one province at the time. Table 2.18 summarizes the range of coefficients for the three outcomes of interest for the preferred model with historical province fixed effects. The coefficients have the expected sign and are statistically significant for all sub-samples suggesting that the findings are not driven from any province in particular.

Randomized Inference: I randomly assign within province the index of dialect similarity. For each outcome, I run 500 regressions of the preferred specification with province fixed effects. Figure 2.5 and Figure 2.6 show the distribution of the coefficients on dialect similarity for the 500 regressions with educational outcomes as a dependent variable. Figure 2.7 displays the distribution of regressions with the incidence of high skilled workers as the dependent variable. The red line represents the point estimate with the real value of dialect similarity. The real coefficient on dialect similarity for the three outcomes of interest lies above the 99^{th} percentile.

2.8 Conclusions

This article exploits linguistic variation within Italy to shed light on whether dialects influence human capital and labour market outcomes. Italian context is suitable to study this link since local dialects differ considerably from Standard Italian. By using unique linguistic data derived from a historical Linguistic Atlas, I find an economically meaningful effect of dialects on both educational and labour market outcomes.

A variety of factors might have shaped dialects and consequently the linguistic similarity to Standard Italian, but I suggest that conditional on historical variables, within province dialects are as good as randomly assigned. By implementing a reduced form model, the central finding of this paper is that conditional on historical, geographical and contemporary controls, current educational outcomes and incidence of the employed in high skilled occupations are positively affected by the linguistic similarity to Standard Italian of dialects prevalent in the municipality at the beginning of the 20th century. The analysis with individual-level data confirms the positive association. An additional finding of this analysis is that linguistic proximity to Standard adversely affects the likelihood of dropout from mandatory schooling.

This paper suggests that dialects may have persistent effects through mainly two possible mechanisms: intergenerational transmission of language skills and intergenerational persistence of educational attainment. I assume that the greater the distance from the national language, the higher the cost to acquire linguistic skills. I test this assumption by employing individual administrative data on standardized Italian test scores of pupils enrolled in the fifth grade of primary schooling. Conditional on preexisting conditions, I find that children living in communities characterized by a historical dialect displaying a greater similarity to the national language perform better in Italian tests. For dialect speaking pupils, the positive effect of linguistic similarity is stronger. Furthermore, I show that language affects academic performance in general; student performance in math scores is positively affected by the extent of dialect similarity to Standard Italian.

Evidence provided in this paper indicates positive association between dialects and economic outcomes, however I am unable to identify a unique mechanism driving the central findings of this paper. Besides, due to internal migration, findings of this paper are likely to be a lower bound. This article focuses on only one aspect of linguistic diversity, i.e., similarity to the standard language. Analyzing the economic consequences of inter-dialectal similarity is an interesting topic for future research.

References

Abedi, J. and Lord, C. (2001). The language factor in mathematics tests. *Applied measurement in education*, 14(3):219–234.

Bian, W., Ji, Y., and Zhang, H. (2019). Does dialect similarity add value to banks? evidence from china. *Journal of Banking and Finance*, 101:226 – 241.

Casey, T. and Dustmann, C. (2008). Intergenerational transmission of language capital and economic outcomes. *Journal of Human Resources*, 43(3):660–687.

Checchi, D., Fiorio, C. V., and Leonardi, M. (2013). Intergenerational persistence of educational attainment in italy. *Economics Letters*, 118(1):229 – 232.

Chiswick, B. R. and Miller, P. W. (2001). A model of destination-language acquisition: Application to male immigrants in canada. *Demography*, 38(3):391–409.

Conley, T. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, 92(1):1 – 45.

De Mauro, T. (1991). *Storia linguistica dell'Italia unita.* Laterza.

Dovì, M.-S. (2019). Does higher language proficiency decrease the probability of unemployment? evidence from china. *China Economic Review*, 54:1 – 11.

Eberhard, D. M., Gary, F. S., and Charles, D. F. (2019). Ethnologue: Languages of the world. twenty-second edition. dallas, texas: Sil international. online version: http://www.ethnologue.com.

Eduard Isphording, I. and Otten, S. (2013). The costs of babylon-linguistic distance in applied economics. *Review of International Economics*, 21:354–369.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2):225 – 239.

Falck, O., Lameli, A., and Ruhose, J. (2018). Cultural biases in migration: Estimating non-monetary migration costs. *Papers in Regional Science*, 97(2):411–438.

Gao, W. and Smyth, R. (2011). Economic returns to speaking 'standard mandarin' among migrants in china's urban labour market. *Economics of Education Review*, 30(2):342 – 352.

Goebl, H. (2008). La dialettometrizzazione integrale dell'ais. presentazione dei primi risultati,. *in: Revue de linguistique romane 72, 25-113.*

Goebl, H. (2016). La geografia linguistica,. *in: LUBELLO; Sergio (a cura di): Manuale di linguistica italiana (Manuals of Romance Linguistics, vol. 13), Berlin / Boston: Walter de Gruyter,553–580.*

Goebl, H. (2018). *Dialectometry*, chapter 7, pages 123–142. John Wiley & Sons, Ltd.

Grogger, J. (2011). Speech patterns and racial wage inequality. *Journal of Human Resources*, 46(1):1–25.

Grogger, J. (2018). Speech and wages. *Journal of Human Resources.*

Guerini, F. (2011). Language policy and ideology in italy. *International journal of the sociology of language*, 2011(210):109–126.

Hsiang, S. M. (2010). Temperatures and cyclones strongly associated with economic production in the caribbean and central america. *Proceedings of the National Academy of Sciences*, 107(35):15367–15372.

Laitin, D. D. and Ramachandran, R. (2016). Language policy and human development. *American Political Science Review*, 110(3):457–480.

Lameli, A., Nitsch, V., Südekum, J., and Wolf, N. (2015). Same same but different: Dialects and trade. *German Economic Review*, 16(3):290–306.

Lepschy, A. L. and Lepschy, G. (1988). *The Italian Language Today.* Routledge.

Maiden, M. and Parry, M. (1997). *The Dialects of Italy.*

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Persyn, D. (2017). On dialects, networks, and labour mobility.

Smith, D. M. (1988). *The Making of Italy, 1796–1866.* Springer.

Suedekum, J. (2018). Economic effects of differences in dialect,. *IZA World of Labor, ISSN 2054-9571, Institute for the Study of Labor (IZA), Bonn, Iss. 414.*

Tosi, A. (2001). *Language and society in a changing Italy*, volume 117. Multilingual matters.

Wang, H., Cheng, Z., and Smyth, R. (2016). Language and consumption. *China Economic Review*, 40:135 – 151.

Wieling, M., Montemagni, S., Nerbonne, J., and Baayen, R. H. (2014). Lexical differences between tuscan dialects and standard italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Yao, Y., Ohinata, A., and van Ours, J. C. (2016). The educational consequences of language proficiency for young children. *Economics of Education Review*, 54:1 – 15.

Yao, Y. and van Ours, J. C. (2018). Daily dialect-speaking and wages among native dutch speakers. *Empirica*.

Yao, Y. and van Ours, J. C. (2019). Dialect speech and wages. *Economics Letters*, 177:35 – 38.

Tables

Table 2.1: Linguistic similarity between some dialects

Pair	*Index of Similarity*
Ploaghe (Sardinia) - Rhemes-Saint-Georges (Aosta)	37.00
Milano (Lombardy) - Milis (Sardinia)	44.19
Bologna (Emilia Romagna) - Palermo (Sicily)	49.00
Standard Italian - Standard French	48.95

Note: The table shows the linguistic similarity between some municipalities (not necessarily the one with the lowest degree of similarity). Regions in parenthesis (Figure 2.8 shows the geographical location of these municipalities). Data source: Goebl (2008)

.

Table 2.2: Dialect similarity

	mean	sd	min	max
Dialect similarity	65.20	7.370	47.09	83.37
Observations	338			

Note: The table shows summary statistics of the index of relative identity value. Data source: Goebl (2008)

.

Table 2.3: Descriptive Statistics of the municipality-level data

	mean	sd	min	max
Mean years of education	9.342	0.718	7.440	11.55
Upper secondary education	34.92	7.571	15.56	60.82
High skilled occupations	26.43	6.452	9.524	50.18
Altitude	2.621	1.542	1	5
Province capital	0.0740	0.262	0	1
Urban/rural	3.142	0.949	1	4
Coastal municipality	0.118	0.323	0	1
Gender ratio	96.14	5.890	83.51	138.2
Demographic density	277.4	828.7	1.649	8082.5
Ageing index	203.5	91.59	76.93	788.2
Latitude	43.38	2.521	37.05	46.70
Longitude	11.61	2.620	6.704	18.30
Latitude (sq)	18.88	2.142	13.73	21.81
Longitude (sq)	1.417	0.632	0.449	3.350
Pop non - urban areas	16.46	15.60	0	74.07
Mean age	45.28	3.096	38.35	57.27
Mean age (sq)	20.59	2.860	14.71	32.79
Distance Florence (log)	5.511	0.753	0	6.692
Distance Turin (log)	5.757	0.900	0	6.978
Literacy (1911)	57.43	26.70	0	97.97
Libraries (1893)	0.133	0.340	0	1
Post office	0.609	0.489	0	1
Railway station	0.127	0.334	0	1
Telegraph office	0.249	0.433	0	1
Port	0.0148	0.121	0	1
Islands	0.112	0.316	0	1
Observations	338			

Note: Table shows descriptive statistics of the municipality level data. Data sources: ISTAT, Italian Census of 1911, Statistica delle Biblioteche (1893), Dizionario dei Comuni del Regno d'Italia (1874). See Table 2.19, Table 2.20, Section 2.4 for a detailed description of the variables and data sources.

Table 2.4: Descriptive Statistics of the province-level data

	mean	sd	min	max
Taxes (1870)	9.054	5.107	0	22.98
Post offices (per 1000 inhabitants in 1869)	0.103	0.0593	0	0.205
Postal money orders value (1869)	7.211	3.676	0	23.30
Number of postal money orders (1869)	91.01	39.63	0	230.3
Enrollment (1871-1872)	43.75	28.15	0	107.4
Province population density (1853-1862)	85.84	83.98	0	781.6
Roads (1871)	112.6	55.37	0	507
Eligible voters (1870)	1.853	0.707	0	4.600
High skilled workers (1871)	0.505	0.183	0	0.859
Journals (1872)	14.85	22.98	0	95
Expenditures per student (1871-1872)	12.27	5.126	0	23.74
Student-teacher ratio (1871-1872)	35.85	12.82	0	64.80
Primary schools(1871-1872)	698.8	562.5	0	2341
Financial Institutions (1872)	0.666	0.472	0	1
Observations	338			

Note: Table shows descriptive statistics of the historical controls at the province level. In parenthesis the year to which variables refer. Data sources: The main source of these data is Annuario Statistico delle provincie italiane per l'anno 1872. High skilled workers 1871 is drawn from the Italian Census of 1871. The source for eligible voters in 1870 is the publication L'Italia Economica nel 1873. See Table 2.21 and Section 2.4 for details.

Table 2.5: Dialects and years of education

	(1)	(2)	(3)	(4)
dep. var.: Mean years of education				
Dialect similarity	0.0234**	0.0159**	0.0151**	0.0263**
	(0.009)	(0.007)	(0.008)	(0.012)
Other Controls	No	Yes	Yes	Yes
Historical Controls	No	No	Yes	Yes
Historical Province Controls	No	No	Yes	No
Province FE	No	No	No	Yes
Historical States FE	Yes	Yes	Yes	Yes
Mean of outcome	9.342	9.342	9.342	9.342
Observations	338	338	338	338
R^2	0.125	0.581	0.648	0.718

Note: The dependent variable is mean years of education. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Historical Province controls include: taxes (1870), post offices (per 1000 inhabitants in 1869), postal money orders value (1869), number of postal money orders (1869), enrollment (1871-1872), province population density (1853-1862), roads (1871), eligible voters (1870), high skilled workers (1871), journals (1872), expenditures per student (1871-1872), student-teacher ratio (1871-1872), primary schools(1871-1872),financial Institutions (1872). Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.6: Dialects and upper secondary schooling

	(1)	(2)	(3)	(4)
dep. var.: Upper secondary education				
Dialect similarity	0.295***	0.192**	0.201**	0.258**
	(0.096)	(0.081)	(0.086)	(0.129)
Other Controls	No	Yes	Yes	Yes
Historical Controls	No	No	Yes	Yes
Historical Province Controls	No	No	Yes	No
Province FE	No	No	No	Yes
Historical States FE	Yes	Yes	Yes	Yes
Mean of outcome	34.92	34.92	34.92	34.92
Observations	338	338	338	338
R^2	0.162	0.537	0.603	0.683

Note: The dependent variable is upper secondary educational attainment rate. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Historical Province controls include: taxes (1870), post offices (per 1000 inhabitants in 1869), postal money orders value (1869), number of postal money orders (1869), enrollment (1871-1872), province population density (1853-1862), roads (1871), eligible voters (1870), high skilled workers (1871), journals (1872), expenditures per student (1871-1872), student-teacher ratio (1871-1872), primary schools(1871-1872),financial Institutions (1872). Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.7: Dialects and labour market

	2011				2001	1991
	(1)	(2)	(3)	(4)	(5)	(6)
dep. var.: High skilled workers						
Dialect similarity	0.161**	0.145**	0.152**	0.265**	0.281**	0.313***
	(0.078)	(0.063)	(0.069)	(0.108)	(0.126)	(0.120)
Other Controls	No	Yes	Yes	Yes	Yes	Yes
Historical Controls	No	No	Yes	Yes	Yes	Yes
Province FE	No	No	No	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE						
Mean of outcome	26.43	26.43	26.43	26.43	32.10	17.79
Observations	338	338	338	338	338	338
R^2	0.0492	0.483	0.546	0.624	0.639	0.728

Note: The dependent variable is high skilled workers. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Historical Province controls include: taxes (1870), post offices (per 1000 inhabitants in 1869), postal money orders value (1869), number of postal money orders (1869), enrollment (1871-1872), province population density (1853-1862), roads (1871), eligible voters (1870), high skilled workers (1871), journals (1872), expenditures per student (1871-1872), student-teacher ratio (1871-1872), primary schools(1871-1872),financial Institutions (1872). Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.8: Nonlinearities: Effects by quartile of dialect similarity

	(1) Mean years of education	(2) Upper secondary education	(3) High skilled workers
1 Quartile	-0.491***	-5.092***	-3.585**
	(0.173)	(1.899)	(1.603)
2 Quartile	-0.413**	-4.311**	-3.482**
	(0.162)	(1.790)	(1.450)
3 Quartile	-0.288**	-3.334**	-2.516**
	(0.136)	(1.490)	(1.224)
Other Controls	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes
Province FE	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes
Mean of outcome	9.342	34.92	26.43
Observations	338	338	338
R^2	0.720	0.686	0.622

Note: Table shows the estimations of the model in equation 2.2 when dialect similarity is expressed as quartiles of its distribution. The fourth quartile is the excluded category. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.9: Dialects and education - LFS individual level data

	(1) Years of education	(2) Upper secondary education	(3) Dropout
Sample of natives			
Dialect similarity	0.0388***	0.00507***	-0.00153**
	(0.013)	(0.002)	(0.001)
Other Controls	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes
Historical Province Controls	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes
Mean of outcome	11.37	0.554	0.0705
No. Cluster	126	126	126
Observations	231544	231544	231544
R^2	0.0893	0.0694	0.0760

Note: Other controls include: age, age(sq), gender, altitude, province capital, urban/rural, coastal municipality, latitude, latitude(sq), longitude, longitude(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Historical Province controls include: taxes (1870), post offices (per 1000 inhabitants in 1869), postal money orders value (1869), number of postal money orders (1869), enrollment (1871-1872), province population density (1853-1862), roads (1871), eligible voters (1870), high skilled workers (1871), journals (1872), expenditures per student (1871-1872), student-teacher ratio (1871-1872), primary schools(1871-1872),financial Institutions (1872). Standard errors in parenthesis (clustered at the municipality level). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.10: Dialects and test scores - Invalsi individual level data

	Italian Score		Math Score	
	(1)	(2)	(3)	(4)
Sample of natives				
Dialect similarity	0.233**	0.189*	0.406***	0.347**
	(0.108)	(0.102)	(0.142)	(0.139)
Dialect similarity x Speaking dialects at home		0.260***		0.340***
		(0.026)		(0.035)
Speaking dialects at home		-0.212***		-0.270***
		(0.019)		(0.025)
Other Controls	Yes	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes	Yes
Mean of outcome	0.788	0.788	0.608	0.608
No. Cluster	4569	4569	4569	4569
Observations	59307	59307	59307	59307
R^2	0.0458	0.0523	0.0354	0.0406

Note: Other controls include: year of birth, year of birth(sq), gender, altitude, province capital, urban/rural, coastal municipality, latitude, latitude(sq), longitude, longitude(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. [a] Dialect similarity and test scores divided by 100. Standard errors in parenthesis (clustered at the class level). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.11: Placebo regressions - LFS

	(1)	(2)	(3)
	Years of education	Upper secondary education	Dropout
Sample of Immigrants			
Dialect similarity	-0.0230	-0.00241	0.000955
	(0.064)	(0.005)	(0.006)
Other Controls	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes
Historical Province Controls	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes
Mean of outcome	9.773	0.424	0.160
No. Cluster	116	116	116
Observations	23376	23376	23376
R^2	0.206	0.0816	0.226

Note: Other controls include: age, age(sq), gender, altitude, province capital, urban/rural, coastal municipality, latitude, latitude(sq), longitude, longitude(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Historical Province controls include: taxes (1870), post offices (per 1000 inhabitants in 1869), postal money orders value (1869), number of postal money orders (1869), enrollment (1871-1872), province population density (1853-1862), roads (1871), eligible voters (1870), high skilled workers (1871), journals (1872), expenditures per student (1871-1872), student-teacher ratio (1871-1872), primary schools(1871-1872),financial Institutions (1872). Standard errors in parenthesis (clustered at the municipality level). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.12: Placebo regressions - Invalsi

	Italian Score	Math Score
	(1)	(2)
Sample of Immigrants		
Dialect similarity	0.0251	0.178
	(0.239)	(0.255)
Other Controls	Yes	Yes
Historical Controls	Yes	Yes
Province FE	Yes	Yes
Historical States FE	Yes	Yes
Mean of outcome	0.699	0.537
No. Cluster	2924	2924
Observations	9823	9823
R^2	0.0907	0.0565

Note: Other controls include: year of birth, year of birth(sq), gender, altitude, province capital, urban/rural, coastal municipality, latitude, latitude(sq), longitude, longitude(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. [a] Dialect similarity and test scores divided by 100. Standard errors in parenthesis (clustered at the class level). Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.13: Mean years of education and linguistic similarity to dialects spoken in big cities

	(1)	(2)	(3)	(4)	(5)
dep. var.: Mean years of education					
Dialect similarity (ref=Turin)	0.0108				
	(0.016)				
Dialect similarity (ref=Milan)		0.000770			
		(0.012)			
Dialect similarity (ref=Venice)			0.0183		
			(0.012)		
Dialect similarity (ref=Bologna)				0.00879	
				(0.012)	
Dialect similarity (ref=Naples)					-0.0150
					(0.012)
Other Controls	Yes	Yes	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes	Yes	Yes
Mean of outcome	9.342	9.342	9.342	9.342	9.342
Observations	338	338	338	338	338
R^2	0.712	0.711	0.714	0.711	0.712

Note: The dependent variable is mean years of education. In each column is shown the coefficient on linguistic similarity from different reference points reported in brackets. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.14: Upper secondary education and linguistic similarity to dialects spoken in big cities

	(1)	(2)	(3)	(4)	(5)
dep. var.: Upper secondary education					
Dialect similarity (ref=Turin)	0.0437				
	(0.195)				
Dialect similarity (ref=Milan)		-0.0368			
		(0.140)			
Dialect similarity (ref=Venice)			0.160		
			(0.139)		
Dialect similarity (ref=Bologna)				0.0376	
				(0.140)	
Dialect similarity (ref=Naples)					-0.178
					(0.140)
Other Controls	Yes	Yes	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes	Yes	Yes
Mean of outcome	34.92	34.92	34.92	34.92	34.92
Observations	338	338	338	338	338
R^2	0.677	0.677	0.679	0.677	0.679

Note: The dependent variable is upper secondary education. In each column is shown the coefficient on linguistic similarity from different reference points reported in brackets. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.15: Labour market and linguistic similarity to dialects spoken in big cities

	(1)	(2)	(3)	(4)	(5)
dep. var.: High skilled workers					
Dialect similarity (ref=Turin)	-0.0319				
	(0.166)				
Dialect similarity (ref=Milan)		-0.00367			
		(0.113)			
Dialect similarity (ref=Venice)			0.0724		
			(0.115)		
Dialect similarity (ref=Bologna)				0.0671	
				(0.119)	
Dialect similarity (ref=Naples)					-0.0327
					(0.125)
Other Controls	Yes	Yes	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes	Yes	Yes
Province FE	Yes	Yes	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes	Yes	Yes
Mean of outcome	26.43	26.43	26.43	26.43	26.43
Observations	338	338	338	338	338
R^2	0.615	0.615	0.616	0.616	0.615

Note: The dependent variable is high skilled workers. In each column is shown the coefficient on linguistic similarity from different reference points reported in brackets. Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.16: Dialect similarity and literacy

	(1)
dep. var.: Literacy(1911)	
Dialect similarity	0.197
	(0.256)
Geographical Controls	Yes
Historical Controls	Yes
Province FE	Yes
Historical States FE	Yes
Mean of outcome	57.43
Observations	338
R^2	0.907

Note: The dependent variable is literacy in 1911. Geographical controls include: latitude, latitude(sq), longitude, longitude(sq), distance Florence(log), distance Turin(log). Historical controls include: libraries(1893), post office, railway station, telegraph office, port. Standard errors in parenthesis. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 2.17: Robustness to omitted variable bias

	(1) Mean years of education	(2) Upper secondary education	(3) High skilled workers
Dialect	0.0263^{**}	0.258^{**}	0.265^{**}
similarity	(0.012)	(0.129)	(0.108)
Beta	0.0381	0.385	0.298
Delta	1	1	1
Rmax	0.933	0.888	0.811
R^2	0.718	0.683	0.624
Observations	338	338	338

Note: Table shows the coefficients of the estimations of the preferred model (equation 2.2) and results of (Oster, 2017) test using Stata command psacalc. Delta refers to the relative degree of selection. *Rmax* is the $R - squared$ of a hypothetical regression which includes observed and unobserved controls. Beta denotes the bias adjusted treatment effect calculated under the assumption that $\delta = 1$, i.e., equal selection and $R_{\max} = 1.3\tilde{R}$, where $\tilde{R}$ is the $R - squared$ from the regressions with controls

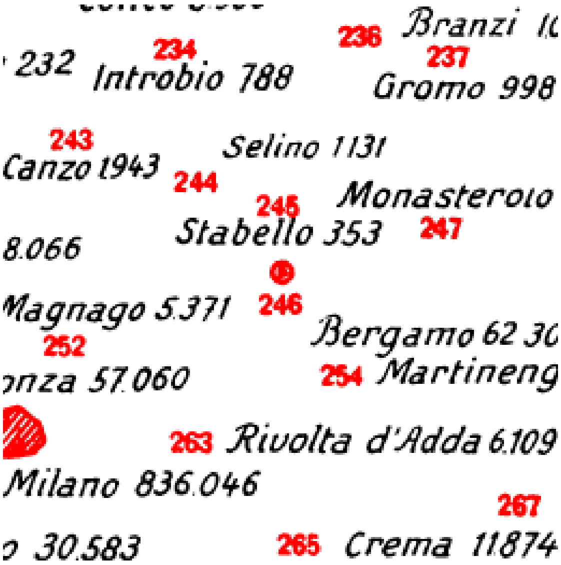
Table 2.18: Robustness checks - range of coefficients subsamples

Outcome	*Coefficients*	*t-statistics*
Mean years of education	[0.021;0.030]	[1.91; 2.54]
Upper secondary education	[0.21;0.30]	[1.68;2.31]
High skilled workers	[0.23;0.33]	[2.13;2.76]

Note: Table shows the range of coefficients and $t - statistics$ of estimations of the preferred model with province fixed effects (equation 2.2) by excluding one province at the time

Figures

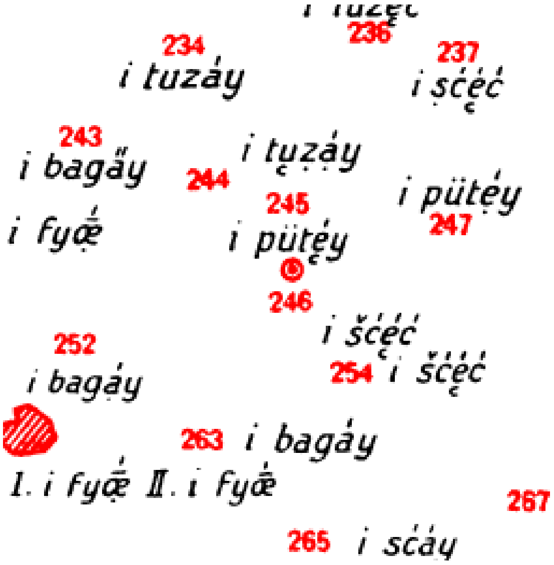
Figure 2.1: Localities (Lombardy)



Source: NavigAIS

Note: Some of the localities surveyed in the region of Lombardy.

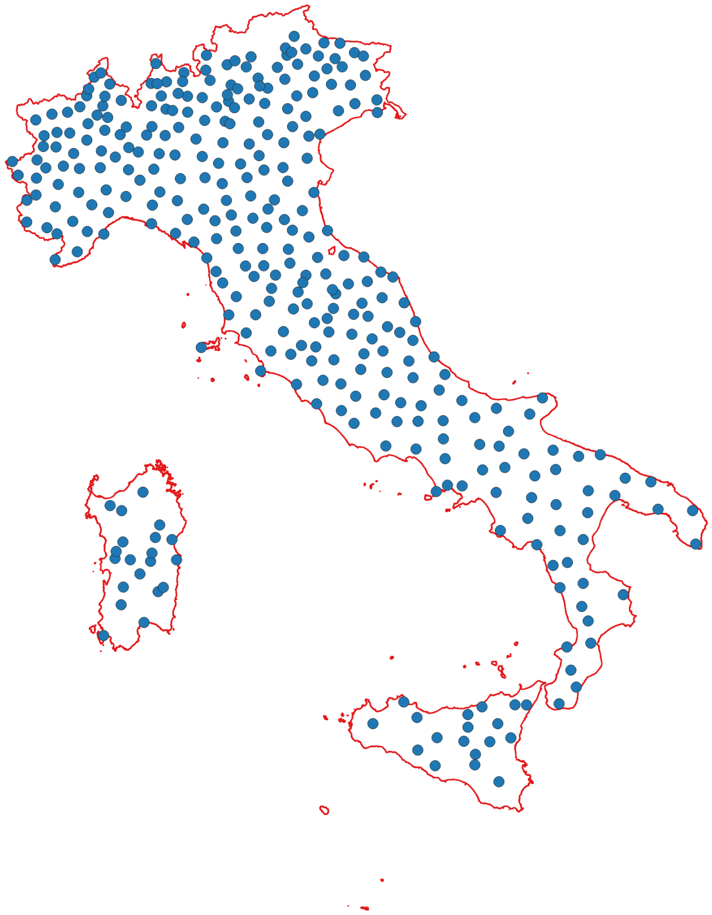
Figure 2.2: The word "I ragazzi"



Source: NavigAIS

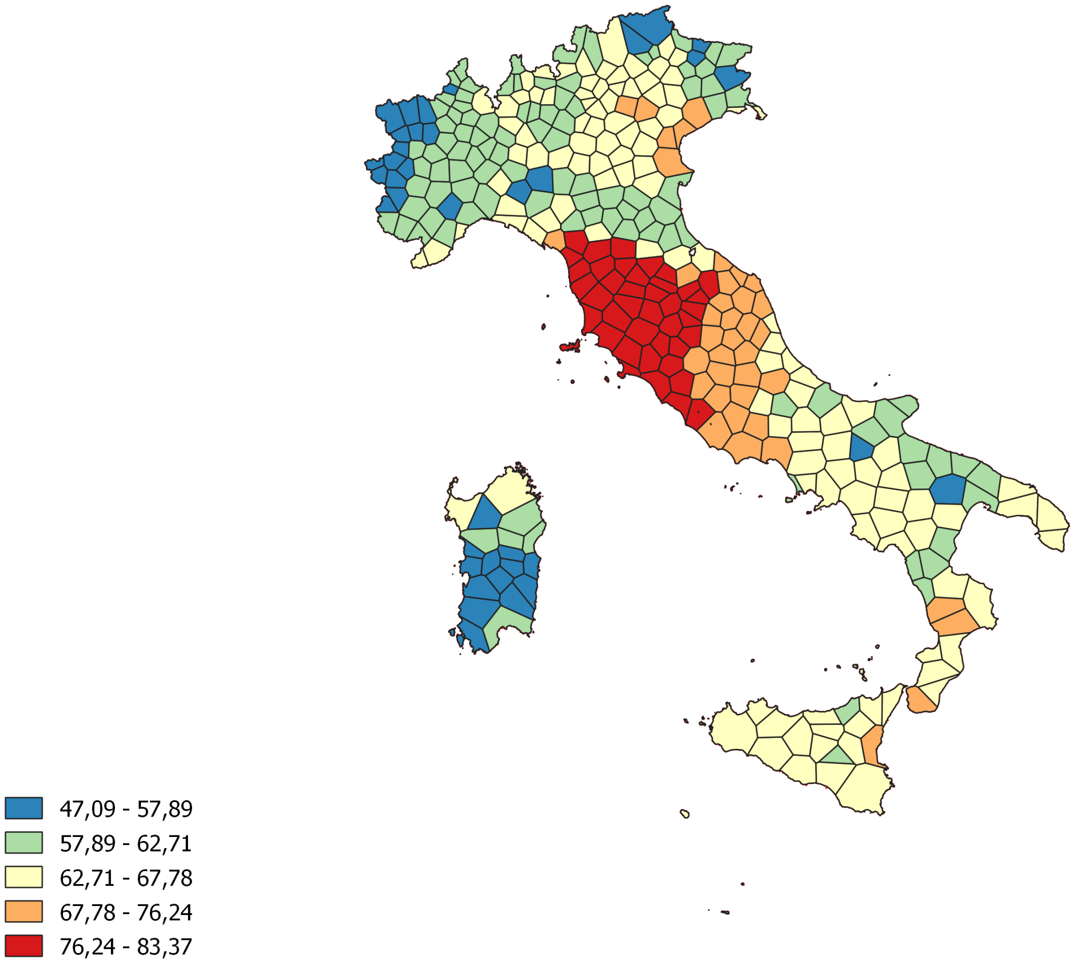
Note: Figure shows the word "i ragazzi" in the localities displayed in Figure 2.1.

Figure 2.3: Localities



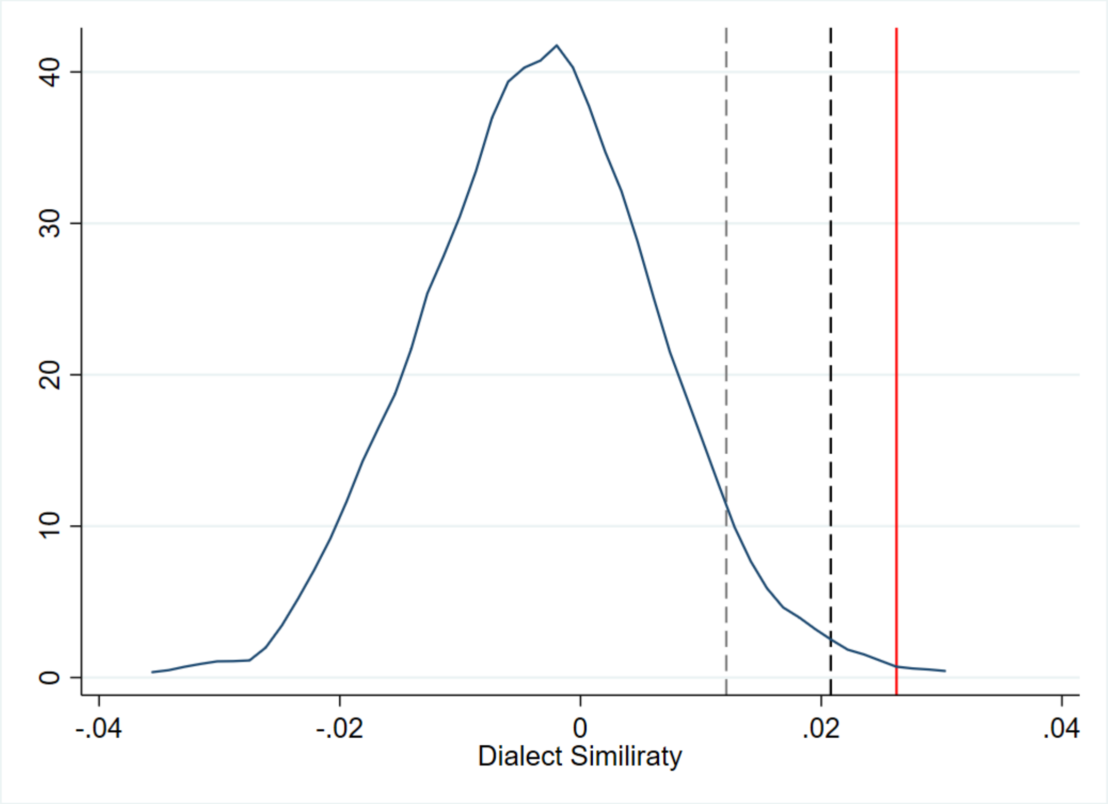
Note: Map of the geographic location of the municipalities of the sample.

Figure 2.4: Map of linguistic similarity to Standard Italian



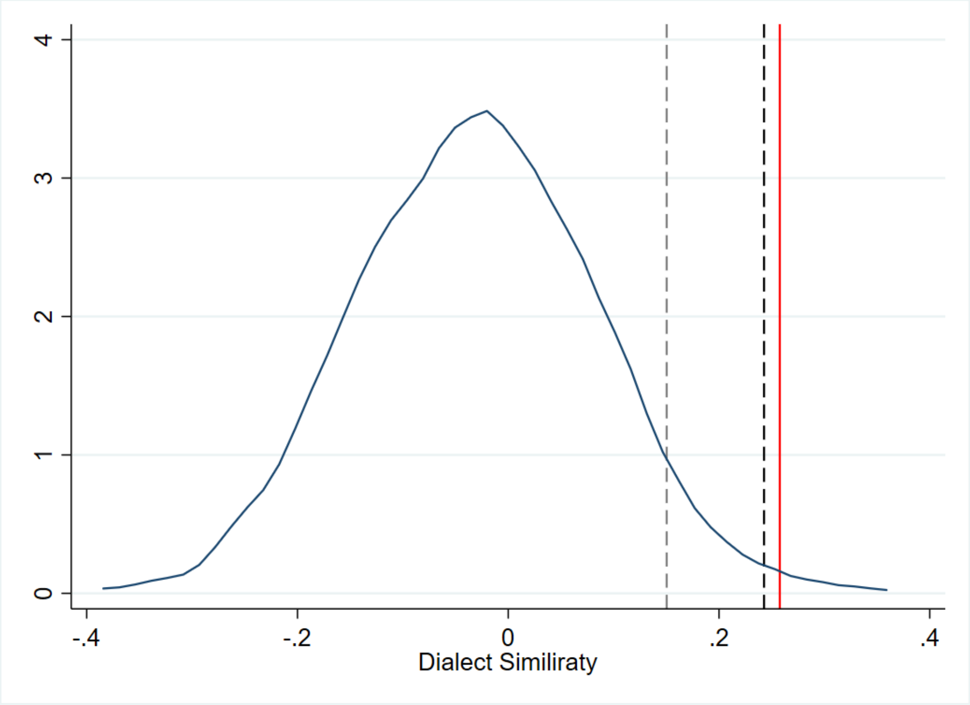
Note: Map of linguistic similarity with reference point Standard Italian. Data source: Goebl (2008)

Figure 2.5: Mean years of education and randomly assigned dialect similarity



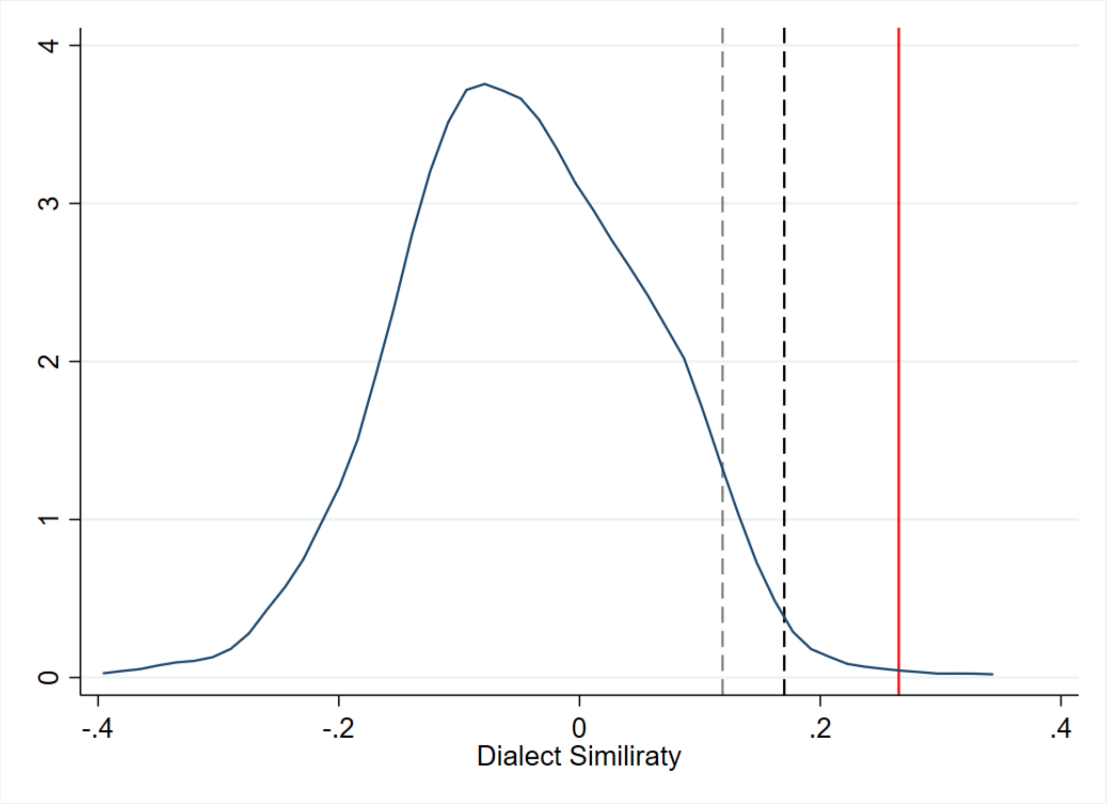
Distribution of the coefficients on dialect similarity on regressions with mean years of education as dependent variable. Dashed lines represents the 95^{th} and 99^{th} respectively. Red lines represent the positions of the estimate based on the real value of dialect similarity.

Figure 2.6: Upper secondary education and randomly assigned dialect similarity



Distribution of the coefficients on dialect similarity on regressions with upper secondary educational rate as dependent variable. Dashed lines represents the 95^{th} and 99^{th} respectively. Red lines represent the positions of the estimate based on the real value of dialect similarity.

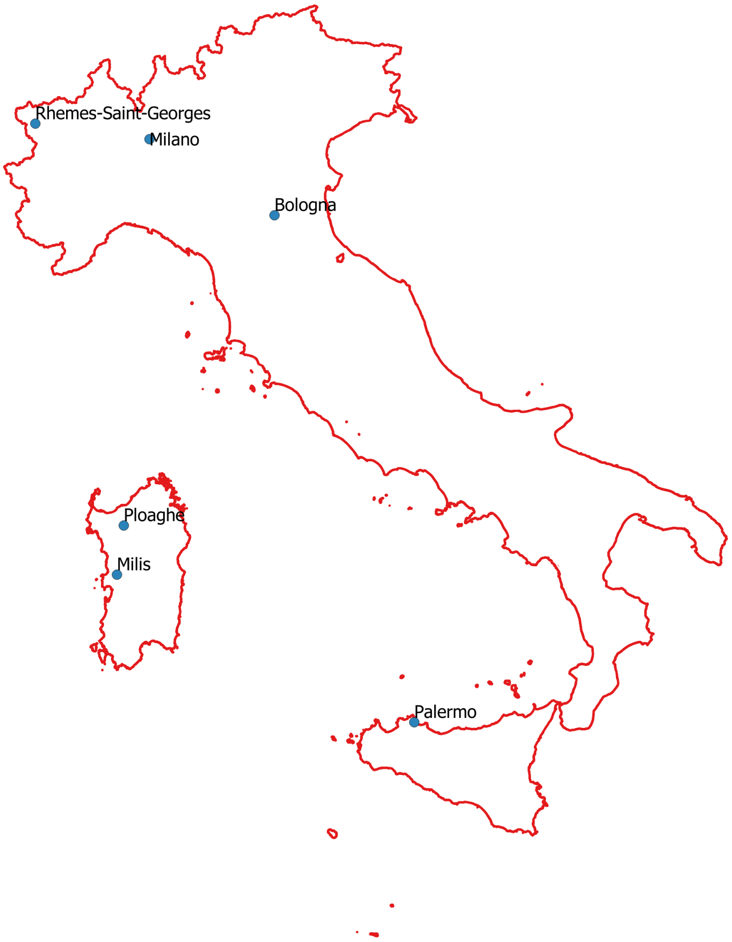
Figure 2.7: High skilled workers and randomly assigned dialect similarity



Distribution of the coefficient on dialect similarity on regressions with the incidence of high skilled workers as dependent variable. Dashed lines represents the 95^{th} and 99^{th} respectively. Red lines represent the positions of the estimate based on the real value of dialect similarity.

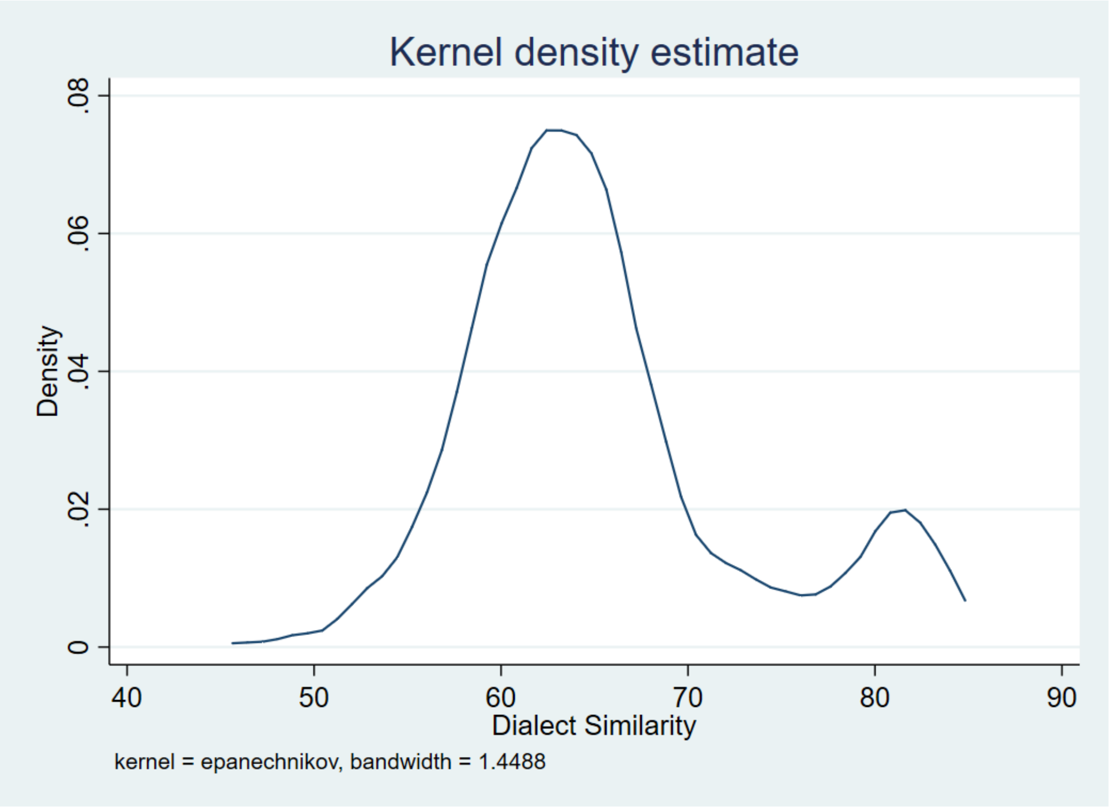
Appendix

Figure 2.8: Geographical location of municipalities in Table 2.1



The figure shows the geographical location of municipalities of Table 2.1.

Figure 2.9: Distribution of the index of relative identity value



The figure shows the distribution of the index of relative identity value.

Table 2.19: Description of contemporary variables

Variable	**Definition**	**Source**
Mean years of education	Mean Years of education of the population aged 18 and over	ISTAT
Upper secondary education	Index of upper secondary educational rate (19 and over)	ISTAT
High skilled workers	Percentage rate of the employed in the following professions: legislators, entrepreneurs and top management, intellectual scientific and highly skilled professions, technical professions	ISTAT
Altitude	Altitude zone (5 categories)	ISTAT
Urban/Rural	Categorical variable (4 categories): rural areas with intensive agriculture, underdeveloped rural areas, intermediate rural areas, urban centers	ISTAT
Province Capital	Binary indicator (=1) if municipality is province capital	ISTAT
Coast	Binary indicator (=1) if coastal municipality	ISTAT
Gender ratio	Ratio of males to females	ISTAT
Demographic density	Resident population per kilometer square	ISTAT
Ageing index	Percentage ratio of the population aged 65 and over to those aged 0-14	ISTAT
Mean age	Mean age of the population	ISTAT
Non - urban population	Resident population living in non-urban areas	ISTAT

2.19 - Continued - Description of contemporary variables

Latitude	Latitude in degrees	ISTAT
Longitude	Longitude in degrees	ISTAT
Islands	(=1) if region Sardinia or Sicily	ISTAT
Distance Florence	log distance from Florence	Based on coordinates from ISTAT
Distance Turin	log distance from Turin	Based on coordinates from ISTAT

Table 2.20: Description of historical variables - municipality level data

Variable	**Definition**	**Source**
Literacy	Percentage rate of population aged 6 years and over who knew to write and read	Italian 1911 Census
Library	(=1) if there was at least a library in the municipality	Statistica delle Biblioteche (1893)
Post office	(=1) if municipality had a post office	Dizionario dei Comuni del Regno d'Italia (1874)
Telegraph office	(=1) if municipality had a telegraph office	Dizionario dei Comuni del Regno d'Italia (1874)
Railway station	(=1) if municipality had a railway station	Dizionario dei Comuni del Regno d'Italia (1874)
Port	(=1) if municipality had a port	Dizionario dei Comuni del Regno d'Italia (1874)

Table 2.21: Description of historical variables - province level data

Variable	**Definition**	**Source**
Taxes (1870)	Direct taxes per inhabitant in 1870	Annuario Statistico delle provincie italiane per l'anno 1872
Post offices (per 1000 inhabitants in 1869)	Number of post offices normalized for the province population in 1871	Annuario Statistico delle provincie italiane per l'anno 1872
Postal money orders value (1869)	Value in lire of postal money orders normalized for the province population in 1871	Annuario Statistico delle provincie italiane per l'anno 1872
Number of postal money orders (1869)	Absolute number of postal money orders normalized for the province population in 1871	Annuario Statistico delle provincie italiane per l'anno 1872
Enrollment (1871-1872)	Percentage enrollment rate in primary schooling of children aged 6-12 years	Annuario Statistico delle provincie italiane per l'anno 1872
Primary schools(1871-1872)	Absolute number of primary schools	Annuario Statistico delle provincie italiane per l'anno 1872
Student-teacher ratio (1871-1872)	Number of students enrolled in primary schooling divided by the number of teachers	Annuario Statistico delle provincie italiane per l'anno 1872
Expenditures per student (1871-1872)	Expenditures per student enrolled in primary schooling	Annuario Statistico delle provincie italiane per l'anno 1872
Roads (1871)	Ratio of length of national and provincial roads to province area	Annuario Statistico delle provincie italiane per l'anno 1872

2.21 - Continued - Description of historical variables - province level data

Variable	**Definition**	**Source**
Eligible voters (1870)	Eligible voters per 100 inhabitants in 1870	L'Italia Economica nel 1873
Province population density (1853-1862)	Population density 1853-1861	Annuario Statistico delle provincie italiane per l'anno 1872
Financial Institutions (1872)	(=1) if province had credit institutions or banks	Annuario Statistico delle provincie italiane per l'anno 1872
Journals (1872)	Absolute number of journals	Annuario Statistico delle provincie italiane per l'anno 1872
High skilled workers (1871)	Rate of population employed in the following categories of professions: law, science and humanities, medicine and education	National Census 1871

Table 2.22: Accounting for spatial correlation

	(1) Mean years of education	(2) Upper secondary education	(3) High skilled workers
Dialect similarity	0.0263	0.258	0.265
OLS	(0.012)**	(0.129)**	(0.108)**
Spatial HAC (50km)	(0.010)**	(0.112)**	(0.085)***
Spatial HAC (150km)	(0.010)***	(0.095)***	(0.072)***
Spatial HAC (250km)	(0.011)**	(0.118)**	(0.081)***
Spatial HAC (350km)	(0.012)**	(0.130)**	(0.101)***
Other Controls	Yes	Yes	Yes
Historical Controls	Yes	Yes	Yes
Province FE	Yes	Yes	Yes
Historical States FE	Yes	Yes	Yes
Observations	338	338	338

Note: Table displays the results of the baseline specification equation 2.2 and standard errors of estimations accounting for spatial correlation under different assumptions on the distance (shown in brackets) at which spatial correlation is assumed to vanish. Conley standard errors are computed using the Stata command *ols*_spatial _HAC provided by (Hsiang, 2010). Other controls include: altitude, province capital, urban/rural, coastal municipality, gender ratio, demographic density, ageing index, latitude, latitude(sq), longitude, longitude(sq), pop non-urban areas, mean age, mean age(sq), distance Florence(log), distance Turin(log), islands. Historical controls include: literacy(1911), libraries(1893), post office, railway station, telegraph office, port. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$