

---

# Linguistic Distance and Job Quality in a Bilingual Labour Market

Lorenzo Cappellari, Antonio Di Paolo and Thompson Ogajah Tawiah

---

---



**Institut de Recerca en Economia  
Aplicada Regional i Pública**  
UNIVERSITAT DE BARCELONA

---

WEBSITE: [www.ub.edu/irea/](http://www.ub.edu/irea/) • CONTACT: [irea@ub.edu](mailto:irea@ub.edu)

---

The Research Institute of Applied Economics (IREA) in Barcelona was founded in 2005, as a research institute in applied economics. Three consolidated research groups make up the institute: AQR, RISK and GiM, and a large number of members are involved in the Institute. IREA focuses on four priority lines of investigation: (i) the quantitative study of regional and urban economic activity and analysis of regional and local economic policies, (ii) study of public economic activity in markets, particularly in the fields of empirical evaluation of privatization, the regulation and competition in the markets of public services using state of industrial economy, (iii) risk analysis in finance and insurance, and (iv) the development of micro and macro econometrics applied for the analysis of economic activity, particularly for quantitative evaluation of public policies.

IREA Working Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. For that reason, IREA Working Papers may not be reproduced or distributed without the written consent of the author. A revised version may be available directly from the author.

Any opinions expressed here are those of the author(s) and not those of IREA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions.

In this paper, we investigate the relationship between language background and labour market outcomes in the bilingual labour market of the Spanish region of Catalonia. The empirical analysis draws on repeated cross-sectional data that allow us to construct a quantitative measure of linguistic distance based on respondents' native language, computed with respect to Catalan, the local language of Catalonia. As labour market outcomes, we consider employment probability and occupational quality, proxied by an indicator for holding a high-skilled job and by an ordinal measure of occupational skill level. The results indicate that, conditional on place of origin and a set of predetermined individual characteristics and controlling for origin-specific trends in years since migration, linguistic distance is not associated with employment. However, it is negatively related to both the likelihood of holding a high-skilled job and occupational skill levels. We analyse the role of language skills as a mechanism, showing that oral and written proficiency in Catalan are key drivers of the negative relationship between linguistic distance and occupational quality. Moreover, this relationship does not appear to be confounded by proficiency in Spanish, and the overall results are robust to a battery of robustness checks. Finally, the analysis of heterogeneous effects reveals an employment penalty associated with linguistic distance among females, and shows that its association with occupational quality is entirely driven by highly educated workers.

**JEL Classification:** J15, J24, J61, Z13

**Keywords:** linguistic distance, employment, occupation, multilingualism

**Authors:**

**Lorenzo Cappellari.** Catholic University of Milan.

**Antonio Di Paolo.** Universitat de Barcelona & AQR-IREA

**Thompson Ogajah Tawiah.** Universitat de Barcelona & AQR-IREA

**Acknowledgements and Funding:**

We acknowledge funding from the Italian Ministry of Research (PRIN grant 2022-J53D23004660008 "STEP-BY-STEP) and from the the Institut d'Estudis de l'Autogovern (grant /23/000004). The usual disclaimers apply.

## 1. Introduction

Individual background is an important driver of socioeconomic outcomes over the life cycle and is intrinsically multidimensional. The economic literature contains a large body of evidence on the relevance of key elements that characterize one's background, such as parental education and occupation, as well as residential location during childhood (Mogstad and Torsvik, 2023). These factors lie outside individuals' control and constitute the "initial conditions" for human capital accumulation and economic performance. As such, individual background accounts for a substantial share of observed inequalities, and its effects are difficult to fully offset through individual investments or public interventions.

Native language constitutes a relevant yet often overlooked dimension of individual background. The language spoken first during childhood is largely determined by parents and may influence future outcomes through multiple channels that operate directly or indirectly via language skills, which are (causally) related to labour market and socioeconomic outcomes. (Aparicio-Fenoll and Di Paolo, 2023). This is especially relevant in the case of migrants, for whom proficiency in the host-country language is crucial for economic and social integration. Existing research shows that migrants' language background, commonly measured by linguistic similarity or distance relative to the destination-country language, significantly affects the acquisition of host-country language proficiency (Isphording and Otten, 2013, 2014). The negative effects of inherited language barriers – measured by linguistic distance – extend beyond limitations on host-country language acquisition. Recent evidence shows that linguistic distance is directly associated with multiple dimensions of immigrants' socioeconomic integration.

One important aspect concerns the effect of linguistic distance on the educational attainment of childhood immigrants. Using cross-country data from the International Adult Literacy Survey (IALS), Isphording (2014) shows that greater linguistic distance between migrants' native language and the destination-country language is associated with lower literacy scores, with effects that increase with age at arrival.<sup>1</sup> Further evidence documents the relevance of linguistic distance for a range of labour market outcomes of adult immigrants across different countries.<sup>2</sup> Using U.S. Census data combined with the

---

<sup>1</sup> Other papers use linguistic distance interacted with age at arrival as an instrumental variable for language proficiency to investigate its causal effect on migrants' test scores using PISA data (Isphording et al., 2016). In this case, however, linguistic distance is defined according to the most common language spoken in the country of origin, given the absence of information on individuals' native language. A slightly different approach is used by Cavallo and Russo (2025) to analyze the causal effect of reading skills on math performance among second-generation immigrants in Italy. Specifically, they use as an instrument for the former variable the interaction between students' age and the linguistic distance between the language spoken at home and Italian.

<sup>2</sup> Beyond labour market outcomes, Bredtmann et al. (2020) find that linguistic distance – defined according to the languages mainly spoken in the home country – also affects the residential location of migrants across different European countries. Moreover, Clarke and Isphording (2017) adopt the same instrumental variable strategy as Isphording et al. (2016) to study the causal effect of language proficiency on the health

O\*NET database, Bacolod and Rangel (2017) investigate the differential effect of linguistic distance by age at migration on occupational sorting, proxied by occupational skill requirements. Adserà and Ferrer (2021), focusing on male workers in Canada, show that immigrants from countries in which the most commonly used language is increasingly distant from English tend to earn lower wages and work in occupations requiring more physical strength and fewer social and analytical skills than natives. Wong (2023) considers the case of asylum seekers in Switzerland, exploiting their allocation to different regions, which generates variation in linguistic distance – given the country’s multilingual reality – while holding the country of origin fixed. The evidence reported in her paper indicates that greater language proximity is associated with higher employment probabilities and a lower risk of earning below the poverty threshold.

However, all these studies investigating the effect of linguistic distance on labour market outcomes share a common limitation: the measure of linguistic distance relative to the destination-country language is defined according to the most commonly used language in the origin country, rather than the individual’s native language.<sup>3</sup> This constrains the interpretation of the resulting estimates as capturing the pure effect of language background, since the proxy is defined at the country-of-origin level, which has two main implications. First, it may capture a variety of country-level features that directly affect labour market performance, such as cultural similarity or institutional factors. Second, it neglects the existence of multilingualism, both at the level of origin countries and within families. The first issue is generally acknowledged by the authors, who typically include various origin-country characteristics as additional controls. Nevertheless, it is difficult to ensure that country-level controls enable fully isolating the effect of linguistic distance from other unobserved country-specific factors that might directly affect the outcome. The second concern, however, has been largely overlooked in the existing literature.

In this paper, we investigate the impact of linguistic distance on labour market outcomes in a bilingual labour market: the Spanish region of Catalonia. Catalonia provides a particularly informative setting for several reasons. First, the region has long attracted substantial migration flows, including internal migrants from other Spanish regions and, since the beginning of the twenty-first century, international migrants from a wide range of origin countries. As a result, the current Catalan population exhibits considerable variation in linguistic backgrounds, which we exploit in the empirical analysis. Second, Catalonia represents a case of asymmetric bilingualism, in which nearly the entire population is proficient in Spanish – with the exception of recently arrived migrants from non-Spanish-speaking countries – whereas only a fraction of residents are fluent in

---

status of immigrants in Australia, again defining linguistic distance based on the most common language in the country of origin.

<sup>3</sup> A related exception is the work by Ghio et al. (2023), in which the authors define linguistic distance from the information about individual’s mother tongue, and use this variable and its interaction with age at arrival (as well as other variables) as instruments for the language proficiency of Italian migrants, in order to investigate the causal effect of language skills on various dimensions of labour market integration.

Catalan.<sup>4</sup> Third, existing evidence shows that proficiency in Catalan – the local language – is rewarded in the labour market not only among foreign-born individuals, but also among internal migrants from other Spanish regions (Rendón, 2007; Di Paolo and Raymond, 2012). Consequently, unlike previous studies on linguistic distance that focus exclusively on international migrants. This represents a key contribution of our paper to the literature. In doing so, we also provide novel evidence on the role of local language skills in multilingual labour market contexts.

We pay special attention in investigating the role of language skills as mechanism, primarily considering oral and written skills in Catalan, as well as the participation in Catalan language courses for migrants, which we take as an alternative proxy for language proficiency. This is also in line with the growing literature on the effects of language training for immigrants' integration (e.g. Lochmann et al., 2019; Foged et al., 2024). When analysing the relevance of language skills as channel, we also consider oral and written competences in Spanish, which is the common language throughout all Spanish regions and its knowledge is quite generalized, even among foreign migrants proceeding from non-Spanish speaking countries. This enables us understanding whether the relationship between linguistic distance and labour market outcomes is driven by the communicative value of language skills. On the contrary, the analyses presented in previous works do not allow separating the communicative value of language skills from other factors that could intervene in the causal chain between language background, language skills and labour market outcomes.

Our empirical analysis draws on repeated cross-section data from a rich survey that includes several language-related variables, as well as detailed sociodemographic and labour market characteristics. The outcomes we examine are employment and occupational status, measured using an indicator capturing whether individuals hold a high skilled and an ordinal scale for occupational skill level. As mentioned above, we focus on individuals born outside Catalonia, either in other Spanish regions (internal migrants) or abroad (international migrants).

Moreover, as previously indicated, we consider self-reported oral and written proficiency in Catalan (and Spanish), along with an indicator for whether the individual has attended a Catalan language course. Most importantly, the survey provides detailed information on respondents' native languages, which we use to construct measures of linguistic distance with respect to Catalan. This allows us to simultaneously control for country-of-birth fixed effects and country-specific trends in years since migration to Catalonia, thereby accounting for differential assimilation patterns across origins. Taken together, these features allow us to overcome key limitations of much of the existing literature on

---

<sup>4</sup> Official statistics for 2023 indicate that among the population aged 15 or older, 99.2% can speak Spanish and 94.5% can write it, while the corresponding figures for Catalan are 80.4% and 65.6%, respectively. (see [https://llengua.gencat.cat/ca/serveis/dades\\_i\\_estudis/dades/sil/poblacio-coneixements-linguistics/](https://llengua.gencat.cat/ca/serveis/dades_i_estudis/dades/sil/poblacio-coneixements-linguistics/)).

the labour market effects of linguistic distance, which typically relies on imputing linguistic distance based solely on country of origin.

The results indicate that, overall, linguistic distance with respect to Catalan is not associated with the probability of employment. However, we find a negative and statistically significant effect of linguistic distance on both the likelihood of holding a high-skilled job and occupational skill levels, highlighting that individuals with a more distant linguistic background tend to hold lower-quality jobs. As expected, the inclusion of place-of-birth fixed effects reduces the magnitude of the linguistic-distance coefficient, suggesting that previous studies relying on country-of-origin imputations may overestimate its true effect. An analysis of potential mechanisms shows that oral and written proficiency in Catalan play a relevant role in mediating the relationship between linguistic distance and occupational selection, whereas we find no clear evidence that attendance at language courses constitutes an important channel. Moreover, the results are not confounded by proficiency in Spanish, which further suggests that the relationship between language background and labour market outcomes operates through channels beyond the purely communicative value of language skills. Overall, the findings are robust to a wide range of robustness checks aimed at validating the interpretation of linguistic distance as a measure of language background. Finally, the analysis of heterogeneous effects reveals the presence of an employment penalty associated with linguistic distance for women, and shows that its negative effect on the probability of having a high skill job and on occupational skill level in general is primarily driven by highly educated individuals, who are more likely to access such occupations.

The remainder of the paper is organized as follows. Section 2 describes the data and presents key descriptive statistics. Section 3 outlines the empirical methodology. Section 4 presents the results, and Section 5 concludes with a discussion of policy implications.

## **2. Data and Descriptive Statistics**

We use data from the Survey of Language Use of the Catalan Population (*Enquesta d'usos lingüístics de la població*, EULP), conducted by the Catalan Statistical Institute (IDESCAT). The EULP is a representative survey of the population of Catalonia aged 15 and older. It provides a unique combination of detailed sociolinguistic information together with standard sociodemographic characteristics and labour market outcomes. Our estimation sample pools the 2013 and 2018 waves, which include 7,255 and 8,722 observations, respectively. For the purposes of this study, we focus on internal migrants born in other Spanish regions (3,050 observations) and foreign-born migrants (2,946 observations). We restrict the sample to individuals aged 18 to 67 (4,525 observations) and exclude 698 observations corresponding to individuals who are students, retired, or permanently

disabled. After dropping a small number of observations with missing values for key variables<sup>5</sup>, the final sample consists of 3,347 individuals.

Our primary variable of interest is the linguistic distance between an individual's native language and Catalan, which we use as a proxy for language background. Accordingly, our main measure relies on information on the respondent's native language – that is, the language first spoken during childhood – to compute linguistic distance. However, taking advantage of the availability of information on both the father's and mother's native languages, we also construct an alternative measure for robustness, defined according to linguistic distance between each parent's native language and Catalan. Following the standard approach in the literature (e.g. Isphording and Otten, 2013, 2014), we employ a quantitative measure of linguistic distance based on lexical similarity derived from the Automatic Similarity Judgment Program (ASJP). The ASJP measure is based on a normalised Levenshtein distance and includes a correction for accidental phonetic similarity between languages, which implies that the resulting distance is not bounded above and can therefore take values larger than 100. For estimation purposes, we use standardised linguistic distance, which has a mean of zero and a variance of one within the estimation sample, to facilitate interpretation of the corresponding coefficients.

Based on data availability, we consider two main dimensions of labour market performance. First, we use employment status to construct a binary variable that takes the value of 1 if an individual is employed and 0 otherwise. Second, we exploit information on the current (or most recent) occupation reported in the EULP survey, coded at the one-digit level and corresponding to the ISCO-08 classification. We therefore convert occupational categories into skill levels, which capture the complexity and range of tasks usually performed in each occupation. The original classification<sup>6</sup> includes four ordinal skill levels. However, due to ambiguity in the classification of managerial jobs, we aggregate the last two categories into a single one, including Managers, Professionals, and Technicians and Associate Professionals.

To explore potential mechanisms, we also pay particular attention to language skills and participation in language courses. Specifically, we exploit self-reported information on oral and written proficiency in Catalan, measured on a 0–10 scale, as well as an indicator equal to 1 if the respondent has ever attended a certified Catalan language course (unfortunately, no information is available on the level attended). Moreover, for robustness checks, we also consider self-reported oral and written skills in Spanish. Finally, our models include as control variables gender, dummies for education and

---

<sup>5</sup> As detailed below, some variables used in the empirical analysis are observed for a smaller number of individuals, either because they are defined only for employed respondents or because of missing values. We do not exclude observations in the latter case, as variables affected by missing values are used only in robustness analyses.

<sup>6</sup> More details can be found in the following link: <https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/>.

parental education, with the latter also containing a separate category for missing values, as well as year of birth, years since migration to Catalonia, and place of birth, which corresponds to the region for individuals born in Spain and to the country for those born abroad.

Table 1 presents summary statistics for the estimation sample. Overall, 67.2% of individuals in the sample are employed. Among them, 28.5% hold high-skill-level jobs, while 16% occupy elementary occupations with the lowest level of skill requirements. Proficiency in Catalan is notably lower than in Spanish: the average self-reported oral skill in Catalan is 4.6 out of 10, and the average written skill is 3.3. By contrast, the averages for oral and written skills in Spanish are 9.1 and 8.6, respectively. Consistently, 61.6% of individuals in the sample are native Spanish speakers (see Table 2), mostly born in other Spanish regions or in Spanish-speaking countries. The average linguistic distance to Catalan is 74.3, with the distance between Spanish and Catalan equal to 72.1. There is, however, substantial variation in linguistic distance across native languages (standard deviation = 16.8), and 12.6% of observations have a lower linguistic distance to Catalan than native Spanish speakers, mostly driven by native speakers of other Romance languages.<sup>7</sup> Moreover, 29% of individuals in the sample have attended a Catalan language course. Regarding sociodemographic characteristics, the sample is fairly balanced by gender, with an average age of approximately 44 years. Internal migrants account for 36% of the sample, while the remaining 64% were born abroad. The average length of stay in Catalonia is around 22 years, substantially higher for internal migrants (35.6 years) than for foreign-born individuals (14.1 years). Parents' place of origin largely coincides with that of the respondents, although a small fraction of individuals have parents born in Catalonia (2.2-2.5%). In terms of educational attainment, 26% of respondents hold a tertiary degree, and 43.3% have completed secondary education. Parental education is lower on average, with only 15.5% of respondents reporting that their parents hold a university degree.

Table 2 disaggregates key variables by linguistic distance groups. The raw data already suggest the existence of differences in labour market outcomes by linguistic distance to Catalan. The employment rate declines with distance, particularly among native speakers of linguistically distant languages, while the share of employed individuals in high-skill jobs is especially high for native Catalan speakers and, to a lesser extent, for native Spanish speakers, and markedly low for individuals whose native language is highly distant from Catalan, who are more likely to hold elementary occupations. Similarly, Catalan language proficiency decreases with linguistic distance. However, the fraction of individuals who have attended a language course is higher among speakers of more

---

<sup>7</sup> More specifically, some individuals in the sample report being native Catalan speakers; these are primarily those originating from other Catalan-speaking areas of Spain (i.e., Valencia, the Balearic Islands, and the northeastern region of Aragón), or individuals with parents born in Catalonia. In addition, the survey includes an option for a mixed Spanish-Catalan native language. For these respondents, we imputed half of the linguistic distance between Spanish and Catalan.

distant languages, possibly reflecting the greater challenges they face in learning Catalan without formal language training. Of course, these differences are unlikely to reflect the causal effect of linguistic background on labour market performance, as they may be confounded by other sociodemographic and cultural characteristics – mostly related to place of origin and differences in assimilation patterns – that correlate with linguistic distance. In the next section, we present the empirical methodology used to isolate the net effect of linguistic distance on labour market outcomes and to examine the role of language skills as a potential mechanism.

### 3. Empirical methodology

We adopt a reduced-form framework that directly relates the (standardised) linguistic distance between the native language of individual  $i$  and Catalan ( $ldist_i$ ) to labour market outcomes. Our starting point is the estimation of the following equation by OLS:

$$Y_i = \alpha + \beta ldist_i + \gamma' X_i + \varepsilon_i \quad (1)$$

As explained above, we consider as outcome variables ( $Y_i$ ) the probability of being employed and, conditional on employment, the probability of holding a high skill job.<sup>8</sup> Moreover, we also model the entire ordinal scale of skill levels using an ordered probit, considering equation (1) as the latent skill requirement associated with different occupations. The vector of control  $X_i$  includes individual characteristics that are correlated with labour market performance and economic assimilation more generally. Specifically, we include gender, year-of-birth dummies, dummies for individual and parental education, and years-since-migration fixed effects. For all estimations, we report standard errors clustered by native language, corresponding to the level of variation in linguistic distance.

However, the OLS estimate of  $\beta$  from equation (1), as well as the corresponding estimates from the ordered probit latent skills equation, are likely to capture not only the effect of language background, but also other factors associated with native language – such as institutional or cultural differences related to the place of origin – that may directly affect the outcome(s). To partial out these unobserved confounders, we exploit our definition of linguistic distance, which is based on native language rather than on the majority language spoken in the place of origin, and augment the specification by including place-of-birth fixed effects ( $\theta_{pb(i)}$ ), which correspond to regions for internal migrants and countries for foreign migrants. The resulting equation consists in:

---

<sup>8</sup> Throughout the paper, when analysing the likelihood of holding a high-skilled job and the ordinal scale of skill level, we focus on the subsample of employed individuals, even though information on occupation refers to the current or most recent job. The results presented below are unaffected by the inclusion of individuals who are currently unemployed and report the occupation in previous employment (available upon request), representing suggestive evidence that self-selection into employment plays a negligible role in our analysis.

$$Y_i = \alpha + \beta \text{ldist}_i + \gamma' X_i + \theta_{pb(i)} + \varepsilon_i \quad (2)$$

In this specification, the coefficient  $\beta$  is identified from within-place-of-origin variation in linguistic distance, exploiting heterogeneity in native languages among individuals from the same country or Spanish region. This variation stems from Spain's multilingual context and indirect migration backgrounds, such as second-generation migrants, as well as from linguistic diversity within country-of-origins and is net of origin-specific institutional and cultural factors absorbed by place-of-birth fixed effects. Moreover, to account for heterogeneity in assimilation by place of origin and differences in migration selectivity, we extend the model to include place-of-birth-specific trends in years since migration, that is:

$$Y_i = \alpha + \beta \text{ldist}_i + \gamma' X_i + \theta_{pb(i)} + \tau_{pb(i)} \text{ysm}_i + \varepsilon_i \quad (3)$$

We retain equation (3) as our preferred specification and interpret the coefficient on linguistic distance in causal terms under the assumption of conditional independence, which results from our ability to control for place-of-birth fixed effects and years-since-migration-specific trends. Nevertheless, to further address potential confounding due to parental characteristics related to the individual's native language – which is largely determined by parental decisions – we conduct three additional robustness checks. First, we estimate an alternative model that additionally controls for the average (standardised) linguistic distance between the parents' native languages and Catalan. Second, we include controls for the father's and mother's (aggregate) place of origin. Third, some individuals may have migrated to Catalonia during childhood, before completing compulsory education. Since 1983, education in Catalonia has been bilingual (see Cappellari and Di Paolo, 2018; Caminal et al., 2021). To account for this, we re-estimate the model excluding individuals who, based on their birth cohort and year of migration, were potentially exposed to Catalan during compulsory schooling.

In a subsequent step, to understand the role of Catalan language skills and language course attendance as potential mechanisms, we replace labour market outcomes with indicators reflecting language proficiency. Specifically, we consider oral and written skills in Catalan, as well as participation in Catalan language courses, which is a strong predictor of language proficiency. To validate the interpretation of these auxiliary regressions as evidence on mechanisms, we also present additional estimates showing the conditional correlation between these three indicators of language skills and labour market outcomes in our sample. However, one could argue that linguistic distance to Catalan is highly correlated with linguistic distance to Spanish, and both languages are relevant in the Catalan labour market, particularly for non-native Spanish speakers. To provide suggestive evidence on the relevance of Catalan skills and to avoid incorrectly attributing the role of mechanisms to local language proficiency, we proceed as follows. First, we report estimates relating linguistic distance to Catalan with oral and written skills in Spanish. Second, we re-estimate the conditional correlations between language

skills and labour market outcomes, while including skills in both languages simultaneously. Significant positive coefficients would challenge the identification of Catalan skills as the relevant mechanism. Third, as further robustness, we repeat our main estimations restricting the sample to individuals with a high level of oral proficiency in Spanish ( $\geq 8$ ).

Finally, we examine the existence of heterogeneous effects of linguistic distance, according to individual characteristics. Specifically, using our preferred specification, we include interaction between linguistic distance and gender, individual's and parental education, (aggregate) place of origins and grouped categories for years since migration.

## 4. Results

### 4.1 *Linguistic distance and labour market outcomes*

Selected estimates of the relationship between linguistic distance to Catalan and labour market outcomes are reported in Table 3. Full results for all outcomes are presented in Tables A1.1-A1.3 of the Appendix. The estimates in the first column correspond to equation (1), which controls only for individual characteristics, including birth-year dummies and years since migration. The coefficient on standardised linguistic distance in the employment equation is imprecisely estimated and not statistically significant. By contrast, a one-standard deviation increase in linguistic distance is associated with a 5.1 percentage point (p.p.) reduction in the probability of holding a high-skilled job, and a decrease of -0.152 points in latent equation for skill level.

In column (2), we include place-of-birth fixed effects. As expected, this leads to a substantial reduction in the coefficients across all outcomes. In particular, the coefficient on linguistic distance becomes very close to zero and remains not statistically significant in the employment equation. As for occupational outcomes, we detect a reduction in the size of the point estimates, although they remain statistically significant at 1% both for the probability of holding a high-skilled job and for the latent skill equation. This pattern confirms the intuition that linguistic distance captures not only language background but also other origin-related characteristics that are relevant for employability, thereby casting doubt on results from existing studies in which linguistic distance is imputed based on country of birth rather than native language, which are possibly overestimating the effect of linguistic distance.

The results obtained after controlling for place-of-birth-specific trends in years since migration (column (3)) are very similar, indicating that the negative impact of linguistic distance on occupational outcomes is not driven by differential assimilation patterns across geographical origins. The estimated coefficients show a reduction of approximately 14% in the probability of holding a high skilled job (mean = 0.28), and of -0.126 points in the latent equation for skill level. Additional computations of average marginal effects from the ordered probit indicate that a standard deviation increase in linguistic distance is associated with a reduction of 2.9 p.p in the probability of having a

job with the highest skill level (3), and an increase of 0.5 p.p and 2.4 p.p in the probability of holding an occupation with a medium skill level (2) and low skill level (1), respectively. Overall, the last specification exhibits a better goodness of fit, according to both the adjusted R-squared and the BIC, and is therefore our preferred model.<sup>9</sup>

#### *4.2 The role of language skills and training*

Table 4 reports the main results from models using indicators of language skills as dependent variables, aimed at investigating their potential role as mechanisms underlying the negative relationship between linguistic distance and occupational outcomes (detailed results are displayed in Tables A4.1-A4.3 of the Appendix). Specifically, we regress self-reported oral and written proficiency in Catalan, as well as an indicator for attendance at a Catalan language course, on the same set of controls used in Table 3, applying the same stepwise inclusion of place-of-birth fixed effects and place-of-birth-specific trends in years since migration. As expected, linguistic distance to Catalan is negatively associated with both oral and written skills. The estimated coefficients are robust in size to the inclusion of fixed effects and trends, and gain statistical precision. In the full and preferred specification, a one-standard deviation increase in linguistic distance reduces oral fluency by 0.49 points and written skills by 0.42 points on the 0–10 scale – quantitatively sizable effects corresponding to 11–13% of the respective averages. By contrast, the probability of having attended language training is unrelated to linguistic distance in any specification.

The evidence reported in Table 4 represents an initial step in interpreting the reduction in language proficiency as a potential channel underlying the negative relationship between linguistic distance and labour market outcomes, particularly regarding occupational quality. Indeed, proficiency in Catalan, the local language of Catalonia, is clearly associated with improved performance, as shown in Table 5. Specifically, the conditional correlation between oral skills in Catalan and the probability of being employed is positive and statistically significant, though quantitatively modest. In contrast, written proficiency and course attendance do not show a significant association with employment. Most importantly, all indicators of language proficiency exhibit a stronger positive correlation with occupational quality among employed individuals – both in terms of skill level and the change of having a high skilled job, supporting the view that language skills serve as a relevant mechanism in shaping labour market outcomes.

To further validate the role of Catalan language skills as a key mechanism, and to address concerns that the previous evidence might partly reflect the relevance of Spanish skills and their potential link to linguistic distance from Catalan, Table 6 presents the results of

---

<sup>9</sup> As shown in Table A2 of the Appendix, the results remain largely unchanged when we include the average (standardized) linguistic distance based on the parents' native languages, as well as dummies for (aggregate) parents' place of birth. Similar evidence is also obtained when excluding individuals who were potentially exposed to Catalan during compulsory education (see Table A3 of the appendix).

the same regressions reported in Table 4, using oral and written Spanish skills as dependent variables. Linguistic distance to Catalan is barely associated with Spanish proficiency, with coefficients much smaller than those for Catalan skills, estimated imprecisely, and not statistically significant. Moreover, Table A5 reports the conditional correlations between language skills and labour market outcomes, simultaneously considering oral and written proficiency in both languages. As shown, only Catalan proficiency exhibits positive and significant coefficients, which remain robust when controlling for Spanish skills, highlighting the importance of the local language in Catalonia's bilingual labour market.

Finally, in Table 7, we repeat our main estimation, restricting the sample to individuals who can be considered orally proficient in Spanish (oral skills  $\geq 8$  out of 10). Once again, the coefficients of interest – especially the negative effect of linguistic distance to Catalan on the probability of a high-skilled job – remain unchanged. The estimate of the latent equation for skill level only suffers a modest reduction, but the qualitative interpretation of the findings is unaffected by this sample restriction. Overall, these results provide further suggestive evidence of the importance of Catalan skills as a mechanism and indicates that our results are not confounded by a potential correlation between linguistic distance and Spanish proficiency. Moreover, these results indicate that the effect of linguistic distance with respect to Catalan on labour market outcomes that we detect is not driven by the communicative value of language skills, since the universal knowledge of Spanish renders Catalan technically redundant from a purely communicative standpoint.

### **4.3 Analysis of heterogeneous effects**

The evidence reported so far indicates that linguistic distance between migrants' native language and Catalan reduces occupational quality, proxied by skill requirements, and decreases the probability of high-skilled job, possibly due to its negative association with language proficiency. However, we do not find any relationship between linguistic distance and employment probability. These aggregate estimates may nevertheless mask heterogeneous effects across key individual characteristics. For this reason, Table 8 presents the results from models that allow for heterogeneous effects by gender, individual and parental education, years since migration, and place of origin.<sup>10</sup> We report marginal effects for each category, along with tests of equality of coefficients across groups.

Gender-specific estimates confirm a negative relationship between linguistic distance and occupational quality, in terms of both skill requirements and the probability of access to high-skilled occupations, with homogeneous effects across genders. In addition, we detect a negative effect of linguistic distance on female employment: a one-standard

---

<sup>10</sup> We also examined heterogeneous effects of linguistic distance on language skills but did not find evidence of differential effects across any of these dimensions (results available upon request).

deviation increase in linguistic distance reduces women's employment probability by 2.5 percentage points. When allowing for differential coefficients by education, we find that the negative association with occupational outcomes is entirely driven by individuals with a university degree, who possess the formal educational credentials required to access high-skilled occupations. As for parental education, although the point estimates of linguistic distance differ across parental education levels, their confidence intervals largely overlap given the associated standard errors, and the p-value of the joint test for coefficient equality clearly fails to reject the corresponding null hypothesis. To analyse heterogeneous effects by length of stay in Catalonia, we group years since migration into three categories: ten years or less, between eleven and twenty years, and more than twenty years. The results show no statistically significant effect of linguistic distance on employment probability across length-of-stay categories. With respect to occupational quality, the negative aggregate effect is driven by individuals who have spent more than ten years in Catalonia. However, the coefficients are imprecisely estimated and, overall, we cannot reject the null hypothesis of equality across categories for any outcome. Finally, we find no differential effects of linguistic distance by place of origin: the negative estimates for both the probability of holding a high-skilled job and latent skill levels are similar – and statistically indistinguishable – between internal and international migrants.

## 5. Conclusion

This paper has studied whether language background, proxied by the linguistic distance between an individual's native language and Catalan, is associated with labour-market outcomes in Catalonia's bilingual labour market. Using two waves (2013 and 2018) of the Survey of Language Use of the Catalan Population (EULP), we exploited a key feature of the data that is typically unavailable in the literature: respondents' native language (rather than an origin-country majority language). This allowed us to construct an individual-level measure of linguistic distance to Catalan and, crucially, to control flexibly for origin-specific factors through place-of-birth fixed effects and origin-specific trends in years since migration.

Three main results emerge. First, once we condition on predetermined characteristics, place of birth, and origin-specific assimilation patterns, linguistic distance is not systematically related to employment in the overall sample. This suggests that, in a context of widespread Spanish proficiency and relatively broad access to employment, language background does not translate into an average employment penalty. At the same time, we do find meaningful heterogeneity: among women, greater linguistic distance is associated with lower employment probabilities, pointing to the possibility that language-related barriers interact with gendered constraints, job search channels, or sectoral sorting.

Second, linguistic distance is robustly and negatively associated with occupational quality. In our preferred specification—controlling for place-of-birth fixed effects and

place-of-birth-specific trends in years since migration – a one standard deviation increase in linguistic distance reduces the probability of holding a high-skilled occupation and shifts workers toward lower occupational skill levels. This pattern is particularly informative because the inclusion of place-of-birth fixed effects substantially attenuates the magnitude of the estimates relative to simpler specifications, indicating that part of what earlier studies attribute to “linguistic distance” (when measured at the origin-country level) may actually reflect broader origin-related cultural, institutional, or selection differences. The remaining within-origin association – identified from variation in native language among people sharing the same country or Spanish region of birth – supports the interpretation of linguistic distance as a meaningful marker of language background in its own right.

Third, the evidence on mechanisms points to local-language proficiency as a central channel linking language background to occupational sorting. Linguistic distance strongly predicts lower self-reported oral and written proficiency in Catalan, with quantitatively sizable effects on a 0–10 scale. In turn, Catalan proficiency is positively correlated with occupational quality among the employed. In contrast, we find little evidence that linguistic distance is related to Spanish proficiency, and the association between Catalan proficiency and occupational outcomes remains when controlling for Spanish skills. This set of results is consistent with the idea that, in Catalonia, Catalan operates less as a purely “communicative” input – given the near-universal availability of Spanish – and more as a locally valued skill that influences access to better jobs, possibly through workplace norms, credentialing practices, customer-facing roles, or the structure of professional networks.

Participation in Catalan language courses does not appear to be systematically related to linguistic distance, despite the strong relationship between linguistic distance and proficiency. This may reflect heterogeneity in course intensity and quality, selection into training based on motivation or constraints, or the limits of measuring training as a simple ever/never indicator (with no information on level or duration). Moreover, the occupational effects of linguistic distance are driven by highly educated individuals. This is an important clue about the margin at which language background matters: linguistic distance seems to be most consequential when individuals are potentially competing for high-skilled positions – where language requirements (formal or informal) and social interaction demands may be more salient – and where Catalan proficiency may serve as a screening device or a complement to other productive skills.

## References

- Adserà, A., & Ferrer, A. (2021). Linguistic proximity and the labour market performance of immigrant men in Canada. *Labour*, 35(1), 1-23.
- Aparicio-Fenoll, A., & Di Paolo, A. (2023). Language Economics. In *Handbook of Labor, Human Resources and Population Economics* (pp. 1-23). Cham: Springer International Publishing.
- Bacolod, M., & Rangel, M. A. (2017). Economic assimilation and skill acquisition: Evidence from the occupational sorting of childhood immigrants. *Demography*, 54, 571-602.
- Bredtmann, J., Nowotny, K., & Otten, S. (2020). Linguistic distance, networks and migrants' regional location choice. *Labour Economics*, 65, 101863.
- Cappellari, L., & Di Paolo, A. (2018). Bilingual schooling and earnings: Evidence from a language-in-education reform. *Economics of Education Review*, 64, 90-101.
- Caminal, R., Cappellari, L., & Di Paolo, A. (2021). Language-in-education, language skills and the intergenerational transmission of language in a bilingual society. *Labour Economics*, 2021, 70, 101975.
- Cavallo, M., & Russo, G. (2025). Lost in Translation: Reading Performance and Math Performance of Second-Generation Immigrant Children in Italy. *Journal of Human Capital*, 19(1), 80-114.
- Clarke, A., & Isphording, I. E. (2017). Language barriers and immigrant health. *Health Economics*, 26(6), 765-778.
- Di Paolo, A., & Raymond, J. L. (2012). Language knowledge and earnings in Catalonia. *Journal of Applied Economics*, 15(1), 89-118.
- Foged, M., Hasager, L., Peri, G., Arendt, J. N., & Bolvig, I. (2024). Language training and refugees' integration. *Review of Economics and Statistics*, 106(4), 1157-1166.
- Ghio, D., Bratti, M., & Bignami, S. (2023). Linguistic barriers to immigrants' labor market integration in Italy. *International Migration Review*, 57(1), 357-394.
- Isphording, I. E. (2014). Disadvantages of linguistic origin—Evidence from immigrant literacy scores. *Economics Letters*, 123(2), 236-239.

Isphording, I. E., & Otten, S. (2013). The costs of babylon—linguistic distance in applied economics. *Review of International Economics*, 21(2), 354-369.

Isphording, I. E., & Otten, S. (2014). Linguistic barriers in the destination language acquisition of immigrants. *Journal of economic Behavior and Organization*, 105, 30-50.

Isphording, I. E., Piopiunik, M., & Rodríguez-Planas, N. (2016). Speaking in numbers: The effect of reading performance on math performance among immigrants. *Economics Letters*, 139, 52-56.

Lochmann, A., Rapoport, H., & Speciale, B. (2019). The effect of language training on immigrants' economic integration: Empirical evidence from France. *European Economic Review*, 113, 265-296.

Mogstad, M., & Torsvik, G. (2023). Family background, neighborhoods, and intergenerational mobility. *Handbook of the Economics of the Family*, 1(1), 327-387.

Rendon, S. (2007). The Catalan premium: language and employment in Catalonia. *Journal of Population Economics*, 20(3), 669-686.

Wong, L. (2023). The effect of linguistic proximity on the labour market outcomes of the asylum population. *Journal of Population Economics*, 36(2), 609-652.

**Table 1: Summary Statistics**

Variable		Obs	Mean	Std. dev.	Min	Max
employed		3,347	0.672	0.470	0	1
<b>occupation (if employed)</b>	<b>skill level</b>					
directors and managers	3	2,210	0.043	0.203	0	1
scientific and intellectual professionals	3	2,210	0.129	0.336	0	1
technicians and support professionals	3	2,210	0.113	0.316	0	1
office, accounting, and administrative clerks	2	2,210	0.047	0.212	0	1
hospitality, personal service, and sales workers	2	2,210	0.260	0.439	0	1
agricultural, livestock, and fishery workers	2	2,210	0.019	0.135	0	1
artisans, industrial, and construction workers	2	2,210	0.136	0.343	0	1
plant and machine operators, and assemblers	2	2,210	0.093	0.290	0	1
elementary occupations	1	2,210	0.159	0.366	0	1
oral skills in Catalan		3,347	4.618	3.461	0	10
written skills in Catalan		3,347	3.269	3.397	0	10
attended a language course		3,347	0.294	0.456	0	1
oral skills in Spanish		3,343	9.136	1.617	0	10
written skills in Spanish		3,343	8.636	2.244	0	10
linguistic distance to Catalan		3,347	74.31	16.76	0	103.25
wave 2018		3,347	0.569	0.495	0	1
female		3,347	0.533	0.499	0	1
age		3,347	44.10	11.52	18	67
born in other Spanish regions		3,347	0.360	0.480	0	1
born outside Spain		3,347	0.640	0.480	0	1
years since migration		3,347	21.83	15.76	1	59
education = primary or less		3,347	0.309	0.462	0	1
education = secondary		3,347	0.433	0.496	0	1
education = tertiary		3,347	0.258	0.438	0	1
parental education = primary or less		3,347	0.573	0.495	0	1
parental education = secondary		3,347	0.238	0.426	0	1
parental education = tertiary		3,347	0.155	0.362	0	1
parental education = missing		3,347	0.034	0.181	0	1
parental linguistic distance to Catalan		3,284	74.44	15.85	0	103.25
mother born in Catalonia		3,340	0.022	0.145	0	1
mother born in other Spanish regions		3,340	0.374	0.484	0	1
mother born in other countries		3,340	0.604	0.489	0	1
father born in Catalonia		3,301	0.025	0.158	0	1
father born in other Spanish regions		3,301	0.376	0.484	0	1
father born in other countries		3,301	0.599	0.490	0	1

Note: figures for occupational categories refer to employed individuals.

**Table 2: Linguistic distances and employment outcomes**

linguistic distance groups:	0 (native Catalan speakers)	55.86-80 (native speakers of Romance languages)	72.12 (native Spanish speakers)	80-95	> 95	Total
%	3.05	13.24	61.61	8.25	13.86	100
% employed	69.61	72.01	70.03	67.75	48.92	67.16
% skill level 3	58.57	24.12	31.12	26.78	10	28.53
% skill level 2	38.57	60.13	53.21	58.47	66.82	55.52
% skill level 1	2.86	15.76	15.67	14.75	23.18	15.95
average oral skills in Catalan	9.12	4.31	4.93	3.05	3.48	4.62
average written skills in Catalan	7.15	3.14	3.30	2.48	2.85	3.27
% attended a language course	22.55	24.15	29.73	29.71	34.27	29.40
average oral skills in Spanish	9.65	8.67	9.77	7.65	7.51	9.14
average written skills in Spanish	9.23	7.86	9.52	6.87	6.36	8.64

Note: The category of native speakers of Romance languages excludes native Spanish speakers, who are classified separately.

**Table 3: linguistic distance and labour market outcomes**

outcome	(1)	(2)	(3)
employment	-0.042 (0.029)	-0.008 (0.010)	-0.007 (0.009)
high skilled job	-0.051*** (0.007)	-0.041*** (0.007)	-0.039*** (0.006)
skill level (latent equation)	-0.152*** (0.030)	-0.138*** (0.024)	-0.126*** (0.026)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes

OLS and ordered probit coefficients of standardized linguistic distance. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries. All models include controls for wave, gender, education and parental education (detailed are reported in the Appendix). Number of observations = 3347 for employment, 2201 for occupation (subsample of employed individuals).

**Table 4: language skills and language training as potential mechanisms**

<b>outcome</b>	<b>(1)</b>	<b>(2)</b>	<b>(3)</b>
oral skills in Catalan	-0.570*** (0.083)	-0.531*** (0.033)	-0.489*** (0.037)
written skills in Catalan	-0.395** (0.160)	-0.450*** (0.061)	-0.423*** (0.057)
attended a language course	0.024* (0.013)	0.008 (0.009)	0.005 (0.009)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes

OLS coefficients of standardized linguistic distance. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries. All models include controls for wave, gender, education and parental education (detailed are reported in the Appendix). Number of observations = 3347.

**Table 5: language skills, language training and labour market outcomes**

	outcome								
	employed			high skilled job			skill level (latent eq.)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
oral skills in Catalan	0.007*** (0.002)			0.018*** (0.002)			0.076*** (0.004)		
written skills in Catalan		0.002 (0.002)			0.021*** (0.002)			0.088*** (0.006)	
attended a language course			0.005 (0.008)			0.048*** (0.013)			0.159*** (0.033)
R-squared/Pseudo R-squared	0.158	0.156	0.156	0.413	0.419	0.404	0.262	0.267	0.250
observations	3347	3347	3347	2201	2201	2201	2201	2201	2201

OLS estimations in columns (1)-(6), ordered probit estimations in columns (7)-(9). Standard errors clustered by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. All regression include fixed effects for year of birth, years since migration and place of birth, years since migration x place of birth trends, and dummies for wave, gender, education and parental education.

**Table 6: linguistic distance with respect to Catalan and skills in Spanish**

<b>outcome</b>	(1)	(2)	(3)
oral skills in Spanish	-0.434*	-0.219*	-0.202
	(0.230)	(0.125)	(0.123)
written skills in Spanish	-0.552*	-0.279	-0.254
	(0.315)	(0.177)	(0.181)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes

OLS estimations, clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries. All models include controls for wave, gender, education and parental education. Number of observations = 3343.

**Table 7: sensitivity to oral proficiency in Spanish (oral skills  $\geq 8$ )**

	outcome					
	employed		high skilled job		skill level	
	(1)	(2)	(3)	(4)	(5)	(6)
estimation sample:	full sample	proficient in Spanish	full sample	proficient in Spanish	full sample	proficient in Spanish
(std) linguistic distance	-0.007 (0.009)	-0.006 (0.007)	-0.039*** (0.006)	-0.037*** (0.006)	-0.126*** (0.026)	-0.108*** (0.023)
R-squared	0.156	0.145	0.407	0.410	0.251	0.262
observations	3347	2939	2201	1984	2201	1984

OLS estimations in columns (1)-(4), ordered probit estimations in columns (5)-(6). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. All regression include fixed effects for year of birth, years since migration and place of birth, years since migration x place of birth trends, and dummies for wave, gender, education and parental education.

**Table 8: Heterogeneous effects**

	(1)	(2)	(3)	(4)	(5)	(6)
	employed		outcome high skilled job		skill level (latent eq.)	
(std) ling. dist. x male	0.010	(0.010)	-0.038***	(0.007)	-0.119***	(0.023)
(std) ling. dist. x female	-0.025**	(0.013)	-0.040***	(0.009)	-0.135***	(0.041)
test for equality of coefficients (p-value)	0.012		0.880		0.687	
(std) ling. dist. x education = primary or less	0.018	(0.017)	0.003	(0.027)	-0.045	(0.078)
(std) ling. dist. x education = secondary	-0.008	(0.009)	-0.022	(0.014)	-0.055	(0.037)
(std) ling. dist. x education = tertiary	-0.028	(0.019)	-0.084***	(0.016)	-0.400***	(0.096)
test for equality of coefficients (p-value)	0.032		0.003		0.001	
(std) ling. dist. x parental education = primary or less	-0.011	(0.011)	-0.033***	(0.008)	-0.118***	(0.039)
(std) ling. dist. x parental education = secondary	0.013	(0.008)	-0.026*	(0.015)	-0.071	(0.054)
(std) ling. dist. x parental education = tertiary	-0.033	(0.180)	-0.072***	(0.049)	-0.267***	(0.097)
test for equality of coefficients (p-value)	0.167		0.429		0.347	
(std) ling. dist. x ( $1 \geq ysm \leq 10$ )	0.010	(0.034)	-0.106	(0.025)	-0.102	(0.086)
(std) ling. dist. x ( $11 \geq ysm \leq 20$ )	-0.009	(0.014)	-0.036***	(0.109)	-0.094**	(0.042)
(std) ling. dist. x ( $ysm > 20$ )	-0.011	(0.007)	-0.044***	(0.006)	-0.146***	(0.024)
test for equality of coefficients (p-value)	0.809		0.431		0.447	
(std) ling. dist. x born in other Spanish regions	-0.006	(0.007)	-0.048***	(0.011)	-0.137***	(0.044)
(std) ling. dist. x born in other countries	-0.009	(0.015)	-0.030**	(0.012)	-0.117***	(0.042)
test for equality of coefficients (p-value)	0.820		0.363		0.766	

OLS estimations in columns (1)-(4), ordered probit estimations in columns (5)-(6). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. All regression include fixed effects for year of birth, years since migration and place of birth, years since migration x place of birth trends, and dummies for wave, gender, education and parental education.

## Appendix

**Table A1.1: linguistic distance and employment (detailed results)**

	(1)	(2)	(3)
constant	0.590*** (0.037)	0.620*** (0.016)	0.620*** (0.017)
(std) linguistic distance	-0.042 (0.029)	-0.008 (0.010)	-0.007 (0.009)
wave 2018	0.101*** (0.015)	0.097*** (0.012)	0.099*** (0.011)
female	-0.143*** (0.018)	-0.158*** (0.021)	-0.155*** (0.021)
education = primary or less	<i>reference category</i>		
education = secondary	0.126*** (0.022)	0.115*** (0.024)	0.112*** (0.023)
education = tertiary	0.233*** (0.029)	0.212*** (0.033)	0.206*** (0.032)
parental education = primary or less	<i>reference category</i>		
parental education = secondary	-0.021 (0.034)	-0.039* (0.023)	-0.041* (0.022)
parental education = tertiary	-0.047 (0.041)	-0.072** (0.030)	-0.067** (0.029)
parental education = missing	-0.054 (0.051)	-0.099** (0.046)	-0.104** (0.050)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
adjusted R-squared	0.085	0.109	0.110
BIC	4091.97	3976.24	3943.70
observations	3347	3347	3347

OLS estimations, dependent variable = employed. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A1.2: linguistic distance and high skilled jobs (detailed results)**

	(1)	(2)	(3)
constant	0.033 (0.022)	0.047** (0.023)	0.048** (0.024)
(std) linguistic distance	-0.051*** (0.007)	-0.041*** (0.007)	-0.039*** (0.006)
wave 2018	0.050*** (0.012)	0.056*** (0.011)	0.057*** (0.012)
female	-0.041*** (0.009)	-0.035*** (0.011)	-0.035*** (0.012)
education = primary or less	<i>reference category</i>		
education = secondary	0.111*** (0.028)	0.101*** (0.026)	0.097*** (0.025)
education = tertiary	0.536*** (0.056)	0.492*** (0.054)	0.480*** (0.054)
parental education = primary or less	<i>reference category</i>		
parental education = secondary	0.030 (0.021)	0.026 (0.020)	0.033 (0.020)
parental education = tertiary	0.137*** (0.021)	0.131*** (0.018)	0.140*** (0.018)
parental education = missing	-0.025 (0.028)	-0.035 (0.030)	-0.038 (0.032)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
adjusted R-squared	0.311	0.351	0.355
BIC	1872.04	1711.95	1666.95
observations	2201	2201	2201

OLS estimations, dependent variable = high skilled occupation (if employed). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A1.3: linguistic distance and skill level (detailed results)**

	(1)	(2)	(3)
constant			
(std) linguistic distance	-0.152*** (0.030)	-0.138*** (0.024)	-0.126*** (0.026)
wave 2018	0.045 (0.039)	0.064 (0.045)	0.072* (0.042)
female	-0.434*** (0.030)	-0.432*** (0.041)	-0.437*** (0.046)
education = primary or less		<i>reference category</i>	
education = secondary	0.412*** (0.086)	0.404*** (0.083)	0.400*** (0.081)
education = tertiary	1.563*** (0.167)	1.501*** (0.173)	1.491*** (0.176)
parental education = primary or less		<i>reference category</i>	
parental education = secondary	0.135 (0.083)	0.120 (0.084)	0.141* (0.082)
parental education = tertiary	0.482*** (0.084)	0.458*** (0.074)	0.486*** (0.076)
parental education = missing	-0.222** (0.103)	-0.294*** (0.110)	-0.302*** (0.112)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
pseudo R-squared	0.196	0.240	0.251
BIC	3774.39	3584.62	3537.28
observations	2201	2201	2201

Ordered probit estimations, dependent variable = skill level (if employed). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A2: parents' languages and origins**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	<b>outcome</b>								
	employed			high skilled job			skill level (latent equation)		
(std) linguistic distance	-0.007 (0.009)	-0.003 (0.011)	-0.017* (0.009)	-0.039*** (0.006)	-0.045*** (0.007)	-0.037*** (0.007)	-0.126*** (0.026)	-0.156*** (0.041)	-0.131*** (0.031)
(std) average parental linguistic distance		-0.007 (0.013)			0.009 (0.010)			0.058 (0.047)	
mother born in Catalonia				<i>reference category</i>					
mother born in other Spanish regions			0.080 (0.051)			-0.024 (0.093)			-0.005 (0.280)
mother born in other countries			0.023 (0.027)			0.014 (0.052)			-0.061 (0.160)
father born in Catalonia				<i>reference category</i>					
father born in other Spanish regions			0.046 (0.031)			0.011 (0.031)			0.122 (0.196)
father born in other countries			0.090** (0.033)			-0.049 (0.039)			0.035 (0.210)
R-squared/pseudo R-squared	0.156	0.159	0.157	0.407	0.409	0.407	0.251	0.254	0.251
observations	3347	3284	3297	2201	2154	2169	2201	2154	2169

OLS estimations in columns (1)-(6), ordered probit estimations in columns (7)-(9). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries. All models include controls for wave, gender, education and parental education.

**Table A3.1: linguistic distance and oral skills in Catalan (detailed results)**

	(1)	(2)	(3)
constant	3.835*** (0.254)	3.792*** (0.171)	3.796*** (0.180)
(std) linguistic distance	-0.570*** (0.083)	-0.531*** (0.033)	-0.489*** (0.037)
wave 2018	-0.483*** (0.125)	-0.461*** (0.124)	-0.464*** (0.137)
female	-0.050 (0.092)	0.012 (0.089)	0.019 (0.084)
education = primary or less	<i>reference category</i>		
education = secondary	1.213*** (0.169)	1.230*** (0.164)	1.215*** (0.156)
education = tertiary	2.074*** (0.242)	1.981*** (0.228)	1.984*** (0.224)
parental education = primary or less	<i>reference category</i>		
parental education = secondary	-0.141 (0.106)	-0.092 (0.096)	-0.090 (0.096)
parental education = tertiary	0.497** (0.206)	0.506*** (0.167)	0.507*** (0.176)
parental education = missing	-0.595** (0.287)	-0.584 (0.364)	-0.580 (0.384)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
adjusted R-squared	0.339	0.355	0.358
BIC	16371.41	16263.09	16216.73
observations	3347	3347	3347

OLS estimations, dependent variable = oral skills in Catalan. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A3.2: linguistic distance and written skills in Catalan (detailed results)**

	(1)	(2)	(3)
constant	2.041*** (0.341)	1.926*** (0.158)	1.939*** (0.161)
(std) linguistic distance	-0.395** (0.160)	-0.450*** (0.061)	-0.423*** (0.057)
wave 2018	-0.273*** (0.072)	-0.254*** (0.069)	-0.268*** (0.072)
female	0.043 (0.074)	0.162** (0.071)	0.163** (0.070)
education = primary or less	<i>reference category</i>		
education = secondary	1.626*** (0.211)	1.675*** (0.182)	1.672*** (0.176)
education = tertiary	2.769*** (0.295)	2.722*** (0.250)	2.709*** (0.231)
parental education = primary or less	<i>reference category</i>		
parental education = secondary	-0.295** (0.109)	-0.217*** (0.080)	-0.219*** (0.078)
parental education = tertiary	0.193 (0.172)	0.263** (0.100)	0.264** (0.108)
parental education = missing	-0.505*** (0.137)	-0.445** (0.168)	-0.448** (0.175)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
adjusted R-squared	0.306	0.327	0.330
BIC	16413.97	16279.28	16238.03
observations	3347	3347	3347

OLS estimations, dependent variable = written skills in Catalan. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A3.3: linguistic distance and language course attendance (detailed results)**

	(1)	(2)	(3)
constant	0.128*** (0.034)	0.116*** (0.028)	0.117*** (0.031)
(std) linguistic distance	0.024* (0.013)	0.008 (0.009)	0.005 (0.009)
wave 2018	0.048*** (0.011)	0.051*** (0.010)	0.050*** (0.011)
female	0.086*** (0.017)	0.094*** (0.015)	0.094*** (0.016)
education = primary or less	<i>reference category</i>		
education = secondary	0.106*** (0.032)	0.110*** (0.032)	0.111*** (0.033)
education = tertiary	0.249*** (0.032)	0.242*** (0.033)	0.246*** (0.037)
parental education = primary or less	<i>reference category</i>		
parental education = secondary	-0.041*** (0.013)	-0.026* (0.016)	-0.027* (0.016)
parental education = tertiary	-0.033 (0.021)	-0.024 (0.019)	-0.028 (0.018)
parental education = missing	-0.073** (0.027)	-0.052* (0.028)	-0.051* (0.029)
year of birth fixed effects	yes	yes	yes
years since migration fixed effects	yes	yes	yes
place of birth fixed effects	no	yes	yes
years since migration x place of birth trends	no	no	yes
adjusted R-squared	0.061	0.071	0.074
BIC	3976.36	3911.58	3870.10
observations	3347	3347	3347

OLS estimations, dependent variable = attended a language course. Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. Place of birth corresponds to the region/country of birth for individuals born in Spain/other countries.

**Table A4: Catalan vs Spanish skills**

	(1)	(2)	(3)	(4)	(5)	(6)
	<b>outcome</b>					
	employed		medium-high skilled job		high skilled job	
oral skills in Catalan	0.007***		0.015***		0.070***	
	(0.002)		(0.002)		(0.005)	
oral skills in Spanish	0.008		0.004		0.014	
	(0.008)		(0.008)		(0.028)	
written skills in Catalan		0.001		0.020***		0.086***
		(0.003)		(0.002)		(0.005)
written skills in Spanish		0.005		-0.002		-0.000
		(0.006)		(0.005)		(0.019)
R-squared	0.173	0.171	0.432	0.438	0.282	0.287
observations	3340	3340	2193	2193	2193	2193

OLS estimations in colums (1)-(4), ordered probit estimations in columns (5)-(6). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. All regression include fixed effects for year of birth, years since migration and place of birth, years since migration x place of birth trends, and dummies for wave, gender, education and parental education.

**Table A5: exposure to Catalan during compulsory education**

	(1)	(2)	(3)	(4)	(5)	(6)
	<b>outcome</b>					
	employed		high skilled job		skill level (latent eq.)	
	full	no Catalan	full	no Catalan	full	no Catalan
estimation sample:	sample	at school	sample	at school	sample	at school
(std) linguistic distance	-0.007	-0.010	-0.039***	-0.033***	-0.126***	-0.102***
	(0.009)	(0.014)	(0.006)	(0.007)	(0.026)	(0.033)
R-squared	0.156	0.161	0.407	0.414	0.251	0.256
observations	3347	2984	2201	1942	2201	1946

OLS estimations in colums (1)-(4), ordered probit estimations in columns (5)-(6). Clustered standard errors by native language in parenthesis, \*\*\* significant at 1%; \*\* significant at 5%; \* significant at 10%. All regression include fixed effects for year of birth, years since migration and place of birth, years since migration x place of birth trends, and dummies for wave, gender, education and parental education.

The logo for UBIREA, featuring the text 'UBIREA' in a bold, white, sans-serif font inside a white rounded rectangle.

## UBIREA

Institut de Recerca en Economia Aplicada Regional i Pública  
*Research Institute of Applied Economics*

**WEBSITE:** [www.ub.edu/irea](http://www.ub.edu/irea) • **CONTACT:** [irea@ub.edu](mailto:irea@ub.edu)

The logo for AQR, featuring a green circular icon with a white dot inside, followed by the text 'AQR' in a bold, white, sans-serif font inside a white rounded rectangle.

## AQR

Grup de Recerca Anàlisi Quantitativa Regional  
*Regional Quantitative Analysis Research Group*

**WEBSITE:** [www.ub.edu/aqr/](http://www.ub.edu/aqr/) • **CONTACT:** [aqr@ub.edu](mailto:aqr@ub.edu)