DISCUSSION PAPER SERIES

IZA DP No. 3061

# Over-Education in Multilingual Economies: Evidence from Catalonia

Maite Blázquez
Sílvio Rendon

September 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

# Over-Education in Multilingual Economies: Evidence from Catalonia

**Maite Blázquez**
*Universidad Autónoma de Madrid*

**Sílvio Rendon**
*Stony Brook University*
*and IZA*

Discussion Paper No. 3061
September 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3061
September 2007

# ABSTRACT

# Over-Education in Multilingual Economies: Evidence from Catalonia*

Catalonia's economy is characterized by linguistic diversity and provides a unique opportunity to measure the incidence of language proficiency on over-education, particularly, whether individuals with deficient language skills tend to acquire more formal skills or, on the contrary, become discouraged to attend school. Descriptive evidence suggests the latter, that individuals with better language knowledge are more likely to be over-educated. However, estimating a model that controls for individuals' socio-demographic characteristics reveals the opposite: better language knowledge decreases over-education. This effect, although robust to accounting for endogeneity of language knowledge and significant at the individual level, is mostly non-significant on average.

JEL Classification: J24, J41, I20, J61, J70

Keywords: over-education, language, immigration, skill premium

Corresponding author:

Sílvio Rendon
Economics Department
Stony Brook University
Stony Brook, NY 11794-0001
USA
E-mail: srendon@ms.cc.sunysb.edu

* We thank participants of the XX Annual Conference of the European Society for Population Economics in Verona, and participants of seminars at U. Pompeu Fabra, U. of Girona, FEDEA, U. Autónoma of Madrid, and U. of Salamanca, as well as Núria Quella and Miguel Sánchez Romero for useful comments. All errors and omissions are only ours.

# 1 Introduction

In a competitive market a variety of qualifications are required, so that individuals who lack one kind of skill tend to acquire more of the others. Specifically, people with deficient language skills may compensate for this disadvantage by accepting jobs for which they have excessive formal skills. This type of skill mismatch may be the case not only for immigrants assimilating to a host economy, but also for natives, since in most countries more than one language is used. This is particularly evident in countries that have undergone language switches, that is, in countries where languages that were once widely but only informally used have become official. In this article we measure the effect of language skills on the quality of job matches in a multilingual economy, an aspect of over-education that has received scarce attention.

An initial descriptive analysis of Catalonia, a multilingual economy, suggests that individuals with better language knowledge appear more likely to be over-educated, rather the opposite of what the substitution-of-skills hypothesis predicts. However, once we estimate a model that controls for several socio-demographic attributes, we find that language knowledge diminishes the probability of over-education. This negative effect, although robust to accounting for endogeneity of language knowledge and significant at the individual level, is mostly non-significant on average.

Over-education occurs when workers' educational qualifications are above those specified for their job or, in other words, when workers are willing to accept jobs with educational requirements below their own. This happens when workers, having failed to find appropriate jobs, refuse to continue their job search. Being a form of labor under-utilization, over-education indicates a poor performance of the labor market. In the literature this phenomenon has been explained as compensation for a lack of human capital, such as experience, ability, or on-the-job training (Sicherman 1991, Alba-Ramírez 1993, Groot 1993, 1996, Groot and Maasen van den Brink 1996). Since language proficiency is a specific form of human capital, over-education may also arise as a consequence of insufficient language skills.

Language proficiency has been shown to improve significantly labor market outcomes, basically earnings, employment probabilities, and unemployment risk (Niesing et al. 1994, Bratsberg and Ragan 1998, Borjas 1999, Schaafsma and Sweetman 2001, Chiswick and Miller 2002, Rendon, 2006). However, studies that relate skill mismatch to language proficiency are few and do not provide clear evidence on how language skills affect the quality of accepted matches.[1] Furthermore, all previous studies deal with this issue in the context of assimilation of immigrants to a host economy. None analyzes the effect of language knowledge on over-education in the context of multilingual economies.

Our main contribution to the literature on over-education is to measure the effect of language proficiency on over-education in a multilingual labor market where language skills potentially play an important role in explaining labor market outcomes. As in other economies characterized by linguistic diversity, public intervention in Catalonia in the form of an active language policy has had direct economic implications. The goal of the language policy carried out by the autonomous Catalan government during the eighties and nineties, called Normalization policy, was to create an economy where the official use of Catalan would increase at the expense of Castilian (Spanish), formerly the only official language. This has obviously contributed to increase the economic value of Catalan knowledge in that individuals with more knowledge of Catalan are significantly more likely to be employed (Rendon 2006). Our work goes a step further and analyzes the incidence of language knowledge on the probability that an individual finds an *appropriate* job. Since the language policy switch affects all workers, whether "immigrants" or "natives," we would expect over-education to be more likely among workers with low Catalan knowledge. Using data for two Census years, 1991 and 1996, we find that once we estimate a model that includes several control variables, such as workers' characteristics, and accounts for endogeneity of Catalan

[1] There is evidence that language fluency increases the likelihood of over-education (Battu & Sloane 2002 for ethnic minorities in Great Britain , and Linsley 2005 for male immigrants in Australia); however, there is also evidence that language skills decrease over-education (Green et al. 2004 for immigrants in Australia, and Barrett et al. 2005 for immigrants in Ireland).

knowledge, the average effect of knowing Catalan on over-education is mostly non significant.

The remainder of this paper is organized as follows. The next section provides a brief overview of the over-education phenomenon. Section 3 gives background to multilingualism and reviews language policy in Catalonia. Section 4 describes the data set and discusses the main descriptive statistics. Section 5 presents the estimation results and Section 6 details the main conclusions of this paper.

# 2 Over-Education as Skill Mismatch

In recent years, the educational level of the labor force in most countries has grown much faster than the supply of high-skilled jobs. This has resulted in a substantial number of workers performing jobs for which they are over-educated. The combined effect of these changes has been a decrease in the demand for lower-educated workers, the ensuing increase in wage inequality, and a widening of unemployment differentials by educational levels, as lower-skilled workers are crowded out by over-educated workers.

As shown by Wieling and Borghans (2001), over-education is labor markets' main adjustment mechanism to an excessive supply of high-skilled workers. In periods of over-supply of high-skilled workers firms substitute away from low-skilled workers, hedging against times when high-skilled workers will be scarce and expensive (Dupuy and de Grip 2002). Moreover, the increased supply of high-skilled labor can reinforce and speed up skill-biased technological change, since firms may complement their high-skilled labor with more use of capital-intensive technologies (Acemoglu 2002).

The concept of over-education was first coined by Thurow (1975) and Freeman (1976) in the context of the U.S. economy of the 1970s.[2] According to Becker (1957),

[2]The coming of age of this now substantial body of literature is reflected in two surveys by Green, McIntosh & Vignoles (1999), by Oosterbeek (2000), Hartog (2000), and an edited book by Borghans & de Grip (2000a) . Both in North America and Europe the literature covers various aspects of imperfect job matching in relation to the educational attainment of workers and the educational

over-education is a temporary phenomenon that occurs during a transition toward a new equilibrium: abundant high-skilled workers earn lower wages and displace low-skilled workers, undermining the incentive for investing in human capital and correcting thereby the initial over-supply of high-skilled workers. Similarly, over-education can be considered an aspect of career mobility, in the sense that it is beneficial for workers to temporarily occupy jobs for which they are over-educated because they gain skills needed to perform higher-level jobs**.** Thus, over-educated workers tend to be younger, change jobs more frequently, and move to better jobs over time either inside or outside the firm (Rosen 1972, Jovanovic 1979, Sicherman and Galor 1990, Sicherman 1991, Robst 1995a, Groot 1993,1996, Groot and Maasen van den Brink 1997, Dekker, de Grip and Heijke 2002).[3] Hence, over-education is a standard feature of a well functioning labor market as part of workers' regular reallocation and qualification process and, therefore, its associated economic costs are negligible. However, other approaches such as signalling (Spence 1973) and job competition (Thurow 1975) theories consider over-education as a long-lasting phenomenon. According to these theories, workers may find it optimal to over-invest in education, in order to signal higher productivity or lower training costs to their prospective employers. Over-education can also be seen as a compensation for the lack of other human capital endowments, such as ability, on-the-job training, and experience.

In the next sections we will describe the process of language switch in a multilingual economy such as Catalonia. Subsequently we will inquire how the over-education phenomenon is intensified by the extension of Catalan knowledge in Catalan labor markets.

---

requirements of jobs.

[3]Recent evidence challenges these findings. Dolton and Vignoles (2000) find that a significant portion (30%) of university graduates are overqualified six years after graduation. Using data from Germany, Büchel and Mertens (2004) Germany find that over-educated workers are less likely to experience upward occupational mobility or above-average wage growth than correctly allocated workers.

# 3 Language Switch in a Multilingual Economy

Practically all countries experience some sort of multilingualism: 66.81% of the world population resides in countries with a language diversity index of 0.459 or more, and 33.29% lives in countries with a diversity index of 0.753 or more. Only 9.95% of the world population lives in countries with a diversity index of 0.134 or less.[4] This implies that in most labor markets, in order to guarantee effective communication individuals have to know more than one language.

Two important language related phenomena occurred on the second half of the twentieth century. On the one hand, the fall of dictatorships in Europe and the process of decolonization in Africa and Asia implied that, in many countries, widely spoken languages that were only informally used gained official status and were given priority in education over the former official, usually colonial, languages. On the other hand, the subsequent waves of immigration to Northern Europe and North America brought the issue of language assimilation of immigrants to their host countries.

Most research on language done by economists has focused on this second issue, approaching language as a form of human capital valued by the market, and crucial to convergence between wages of immigrants and natives. A recent stream of the economic literature studies changes in the language of education and attempts to measure their economic effects. As examples we have the language switch from French to Arabic in Morocco, from English to Welsh in Wales, from Russian to Estonian in Estonia, or from English to Spanish in Puerto Rico (see Angrist and Lavy 1977, Grin and Vaillancourt 1998, Sabourin and Bernier 2003, Angrist et al. 2006). However**,** evidence on the economic effects of switching from a metropolitan to a local language

[4]These numbers are computed by the authors based on Table 6 of Gordon (2005) . Language diversity in a country is measured by Greenberg's diversity index, defined as $D = 1 - \sum_{k=1}^{K} p_k^2$, where $p_k$ is the fraction of total population speaking language $k$, and $K$ is the total number of languages spoken in a country. This index represents the probability that any two randomly selected individuals of the country have different mother languages. For total diversity and no two individuals having the same mother language this index is 1; for no diversity at all and everyone having the same mother language this index is 0.

is so far inconclusive. While Angrist and Lavy (1977) find that the language switch in Morocco decreased returns to education, Angrist et al. (2006) show that once education-specific cohort trends were introduced, English instruction had no effect on English-speaking ability among Puerto Rican natives.

Similarly to other European countries, Spain is characterized by a vast language diversity. Altogether, around forty percent of the population of Spain lives in areas with two official languages. While Castilian (Spanish) is official in the totality of the territory, Catalan, Galician, and Basque share co-officiality with Castilian in their own territories.[5]

During most of its history the Catalan language has been official in Catalonia. However, from the forties to the seventies, during Franco's regime, Spanish was declared the only official language in Catalonia (and actually in the whole of Spain), while Catalan was reserved for private use. This repression, combined with massive immigration of Spanish speakers to Catalonia, helps to explain how an important proportion of Catalans did in the recent past not master Catalan, even if it was their native language.

After Franco's regime, the Catalan language regained officiality; the reestablished Catalan autonomous government provided several incentives for its propagation and wider use as the normal language of public and private communication. The "Normalization policy" (*normalització*), enacted in the early eighties, extended the use of Catalan in the fields of education, public administration and public media. In this period Catalan progressively replaced Castilian as the main language of instruction in primary, middle, and secondary schools.

Figure 1 shows the evolution of spoken Catalan in the three main Catalan-speaking

[5] Catalan is official in Catalonia (6,995,206 inhabitants in 2005), in Valencian Country (4,692,449 inhabitants), and in Balearic Islands (983,131). Galician is official in Galicia (2,762.198) and Basque is official in the Basque Country (2,124,846) and in the north of Navarra. These regions represent 39.81% of the total Spanish population (44,108,530). Other languages are Asturian or Bable (with around 600,000 speakers and not official in its territory: Asturias and the north of Castilla-Leon), and Aranés (Occitan, official in Vall d'Aran, within Catalonia, with 9,100 inhabitants).

areas. In Catalonia, the only of the three with active public policies to foster language knowledge, there is a recovery of spoken Catalan from the mid-seventies onwards. In Valencia and Balearic areas, Catalan is also co-official with Castilian, but public policies have been less active in promoting its knowledge and use.

Normalization policy has been especially successful in those sectors in which the Generalitat had exclusive competencies. In contrast, success has been less prominent in those which were not regulated at the time, such as the private and socio-economic domains[6]. Unlike in the Basque country, where reading, speaking, and writing skills increase in similar proportions, in Catalonia writing skills increase the most (Casesnoves, Turell and Sankoff 2006). Moreover, the increased use and knowledge of Catalan does not undermine the use of Castilian. In particular, Castilian-speakers have increased their knowledge of Catalan at the same time as they have maintained their intensive use of Castilian (Vila and Vial 2001). And interestingly, as remarked by Woolard (1989), Castilian speakers born outside Catalonia are more likely than those born in Catalonia to attempt to learn Catalan. Despite these facts, since 1983 there has been a steady increase in the knowledge of Catalan in the whole of Catalonia, in all spheres and among all age groups. In the next section, we will provide descriptive evidence for over-education rates by level of Catalan knowledge and several demographic characteristics.

# 4 Data

We use two samples of 250,000 randomly selected individuals, extracted from census data for 1991 and 1996,[7] provided by the Catalan and Spanish National Statistical Institutes (IDESCAT-INE). These datasets contain information on personal attributes such as gender, age, marital status, schooling, place of residence, place of birth, num-

[6]Figures for the 2001 census have not yet been made available to the general public.

[7]Unlike the census of 1991, applied in all of Spain, the census of 1996 was only applied in Catalonia.

ber of years in Catalonia, occupational status, and knowledge of Catalan. We combine this information with data at the district area, called *municipi*, to capture the externality effects of residing in areas with high employment rates and/or widespread Catalan knowledge.

We restrict the sample to only parents and children aged between 16 and 60, born in Spain and participating in the labor force. The final sample contains 96,863 individuals for year 1991, and 96,985 individuals for 1996. Appendix A.1 details the sample selection.

We consider an individual as over-educated when he/she has more years of schooling than the mean for his/her occupation plus one standard deviation. Appendix A.2. describes in greater detail the definition of over-education and under-education as well as the rest of variables.

Knowledge of Catalan, classified into understanding, reading, speaking and writing, is self-reported. Because Catalan is linguistically close to Castilian, respondents may over-report their knowledge of Catalan. To alleviate the possible biases caused by self-reporting and linguistic closeness (Charette and Meng 1994), we class individuals who claim to either understand, only speak, or only read Catalan as having a basic level of Catalan knowledge; individuals who report to read *and* speak will be in the intermediate level, while those who can write have a superior level of Catalan knowledge.

Descriptive statistics for all variables by gender and census year are presented in Table 1, where we can see that over- and under-education each amount to between 9% and 15% of the labor force, that is, the skill mismatch involves in between 18% and 30% of the labor force. Gender differences in over-education rates are not systematic: while in 1991 over-education is less likely among females, the opposite is observed in 1996. However, the incidence of over- and under-education decreases over time for both genders. This improvement in skill matching is associated with an increase of the educational attainment of the whole population in this period. Average years of

schooling of adequately educated workers increased from 7.7 and 8.6 to 8.5 and 9.3 for men and women, respectively. We also observe that over-educated workers of both genders tend to be younger than adequately educated workers, while the opposite is true for under-educated workers. Therefore, skill mismatch mainly affects young workers and tends to diminish over time as workers' schooling increases.

The descriptive analysis clearly suggests a positive relationship between over-education and Catalan knowledge. It is more likely to find individuals who read and speak, and write Catalan among the over-educated than among the adequately educated. On the contrary, among the under-educated it is less likely to find individuals who read and speak, and write Catalan. Conversely, the probability of being over-educated is higher among individuals, male or female, who exhibit an intermediate or superior level of Catalan knowledge. Notice also that average knowledge of Catalan increases over time. Furthermore, women are always more proficient than men.

The proportion of over-educated individuals in the service sector is above the average for the overall economy, although it decreases over time. Moreover, it is worth noting that the percentage of female workers in the service sector is also considerably higher than the economy's average. As we will see below, this difference may be the underling reason for the over-education gender gap. Interestingly, average marriage rates are higher for men than for women. The percentage of the population directly affected by the Normalization process is higher for women than for men and it is growing over time for both genders: from 3.3% in 1991 to 10.2% in 1996 for men, and from 3.0% in 1991 to 12.7% in 1996 for women.

The population of Catalonia is strongly concentrated: 80% reside in the province of Barcelona, although this percentage is decreasing over time. Girona, Tarragona and Lleida, in this order, follow in importance as provinces of residence. Data at the municipal level confirm the population's increasing Catalan proficiency, as well as an increase in the service sector's share in the economy and a decrease of employment

rates for both genders.

Approximately one third of the sample consists of people born outside Catalonia; within this group, more than two thirds come from Andalusia alone, while very few come from other Catalan-speaking areas such as Valencia, the Balearics or La Franja (a thin stretch of land in neighboring Aragon). The data also reveal a reduction of the proportion of individuals born outside Catalonia and of the proportion of men and women born in Andalusia. Individuals born outside Catalonia arrived mostly at the end of the sixties and have been in Catalonia for an average of between 23 and 25 years in 1991, and between 27 and 28 years in 1996. Furthermore, around one third of them arrived when they were no older than 10.

In sum, the descriptive analysis suggests that the knowledge of Catalan increases the probability of over-education and decreases the probability of under-education. This first approximation can lead us to think, as it is usually the case in the public discussion on language policy in Catalonia, that individuals who are not fluent in Catalan tend to be under-educated, that is, they are discouraged to attend school because of the switch in the language of instruction. In the next section, we will see how this picture changes once we control for several socio-demographic attributes and account for endogeneity of Catalan knowledge.

# 5 Estimation

In this section we proceed to a more in-depth analysis of how over-education in Catalonia is related to Catalan knowledge. To this purpose, we first perform probit estimations for years 1991 and 1996, for males and females separately. As explanatory variables we include personal characteristics such as schooling and its square, age and its square, an interaction term between schooling and age, marital status, and a set of occupational and sectorial variables. Furthermore, we also control for local labor market characteristics by including as explanatory factors the employ-

ment rate in the *municipi*, the share of those employed in the service sector in the *municipi*, and a set of regional dummies. Finally, alternative estimations are done for the intermediate (reading and speaking) and superior (writing) levels of Catalan knowledge. Under this standard probit approach Catalan knowledge is considered as an exogenous explanatory factor.

## 5.1 Over-Education by Level of Language Knowledge

Table 2 shows the representative and average discrete effects and Catalan premia obtained from standard probit models for over-education, presented separately for the intermediate and superior level of Catalan knowledge. Discrete effects are defined as the variation in the probability of over-education produced by a discrete variation in the level of Catalan knowledge. *A representative individual's discrete effects* are called representative discrete effects, whereas average discrete effects correspond to *the average of discrete effects over all individuals.*

This table shows that representative discrete effects of Catalan knowledge on the individual likelihood of being over-educated are negative and significant for almost all groups, although the magnitude of the effect in absolute terms is smaller when occupational and activity dummies are included in the analysis. Average discrete effects are also found to be negative and with greater absolute values, but, unlike representative discrete effects, they are clearly not significant. For example, for a woman of average socio-demographic characteristics, reading and speaking Catalan in 1996 decreases the probability of her being over-educated by 0.29 percentage points and this effect is significantly different from zero. However, for all women in 1996 reading and speaking Catalan decreased the probability of being over-educated on average by 3.16 percentage points, and this contribution is non significantly different from zero. The difference between representative and average effects is produced by the curvature in the cumulative function (Jensen's inequality). In contrast, the higher standard deviations for average effects are caused by the strong variation of

effects across individuals, which is not considered in the standard deviation of effects for a representative individual. The net result of higher means and higher standard deviations for average effects is that the average discrete effects are non significant.

Both representative and average effects of language knowledge on over-education are larger for women than for men. Moreover, these effects are decreasing over time, as they fall from 1991 to 1996, and they are decreasing over the level of proficiency in the language, as the effects are smaller for writing than for reading Catalan.

It is interesting to notice how the results obtained from the standard probit analysis clearly differ from what the descriptive statistics suggested. As we recall from Table 1, the percentage of over-educated workers is significantly higher among those with an intermediate or superior level of Catalan knowledge. This is observed both in 1991 and 1996 and for both males and females. In 1996, for instance, 14.3% of females with superior level of Catalan knowledge were over-educated, while the corresponding percentage among those with no Catalan knowledge was 5.4%. However, the estimation results obtained from a standard probit reveal that a superior level of Catalan knowledge significantly diminishes the individual likelihood of over-education.

These results suggest the existence of some degree of substitution between language skills and educational attainment: workers with Catalan knowledge are more likely to occupy jobs where their educational attainment match better with the jobs' educational requirements. The opposite is true for workers without Catalan knowledge, who have less opportunities to work in jobs that suit their education level and, therefore, have diminished career prospects. Notice, however, that this negative effect of Catalan knowledge on over-education is only significant at the individual level. On average, because of the strong variation across individuals, one cannot reject the hypothesis that this effect is zero.

## 5.2 Endogeneity of Language Knowledge

The previous estimations constitute an important evidence of the effect of Catalan knowledge on over-education. They provide unbiased results if language knowledge is an exogenous variable, i.e. if language were merely an ethnic attribute that signals membership to a given community, as assumed by the early studies on the economics of language (Becker 1957 and Reynauld and Marion 1972; see also Grin 2003). These estimations, however, do not account for the possible endogeneity of language knowledge, that is, the effect estimated by a standard probit may be driven by attributes that account for both language knowledge and over-education. Even in this non-linear limited-dependent variable estimation, the effect of the unaccounted language variables on over-education could contain the effect of other variables that determine Catalan knowledge. Hence, if Catalan knowledge is not exogenous the estimation results of a standard probit will be biased. Thus, in order to account for selection into knowing Catalan, we estimate the probability of being over-educated conditional on the level of Catalan knowledge by gender and Census year.

We proceed in a two-step estimation, as in Willis and Rosen (1979). In the first stage, estimation selection into Catalan knowledge is accounted for by the same variables of the over-education equation, explained above, augmented by the following: percentage of individuals born in Catalonia and percentage of individuals who write Catalan in the *municipi*; a dummy variable indicating whether the individual was affected by the Normalization process; whether the individual arrived to Catalonia before age 10; number of years since migration; an interaction term between years since migration and whether the individual arrived before age 10; and, finally, origin variables such as whether the individual was born in Andalusia, Valencia and Balearics, or La Franja.[8] As exogenous sources of variation, these variables allow us to identify the recursive bivariate probit model (Maddala 1983); thus, they are

[8]The estimation results of the first stage are very similar to those obtained in Rendon (2006). The details of the estimation are available upon request.

only included in the language selection equation, but excluded in the over-education equations. The variables assumed to affect Catalan knowledge but not directly the probability of over-education represent the externality effect of the community of residence on Catalan knowledge, the exposure to schooling in Catalan, and the language predominant in the environment where the individual was raised.

Introducing origin variables as instruments may at first seem objectionable, partly because individuals who were born outside their current place of residence are usually considered as immigrants and, therefore, different from natives in more than their knowledge of the local language. However, one has to bear in mind that in this particular context we are comparing Spanish citizens perfectly equivalent in physical (racial), religious, and legal terms. Moreover, individuals born outside Catalonia are not unassimilated newcomers; as seen above, they arrived in Catalonia when they were, on average, between 23 and 25 years old and one third of them before the age of 10. As such, and irrespective of their level of knowledge of the Catalan language, they identify themselves as Catalan and are considered Catalan by those born in Catalonia. Our case of study has the additional specificity that, unlike in most economies, in Catalonia not all natives know their language, but all of them know the language of the individuals born outside, Castilian. These particularities substantiate origin variables as instruments that are only correlated with Catalan knowledge, but not with over-education.[9]

Using the estimated parameters of the language selection equation, we proceed to estimate the second stage: the probability of being over-educated conditional on a given Catalan proficiency level. Table 3 presents the predicted probabilities of being over-educated by Catalan reading and speaking, and writing skills based on previous estimations. As shown in the descriptive statistics, for all groups the actual probability of being over-educated is substantially higher for individuals who know

[9]Nevertheless, results based on this identification strategy should be taken cautiously, as the independence assumption and thereby the identification of the language premium weakens if natives have better networks and end up being better placed than individuals born outside Catalonia. Thus, these results can be improved when richer datasets and wider sets of instruments become available.

Catalan. However, this comparison does not reveal much, because it is made across groups with different individual attributes. A valid comparison of possible outcomes should be made for the same subgroup of the population, and that is possible after recovering the parameters of the language and the over-education equations. This is contained in the average discrete effects of Catalan knowledge, which are found to be negative, suggesting that Catalan knowledge indeed reduces the likelihood of over-education. For instance, the illusion that Catalan knowledge increases over-education may be a result of naively comparing the probabilities of over-education for individuals who know and for individuals who do not know Catalan, 17.62% and 8.93%, respectively. In contrast, Catalan knowledge decreases over-education by 1.85 percentage points for individuals who know Catalan and by 0.84 percentage points for individuals who do not know Catalan. However, the standard deviation of these effects is so high that they end up being not significant: one cannot reject that all these effects are significantly different from zero.

Note that the effect of language on over-education is higher for men than for women, decreasing over the level of Catalan proficiency, decreasing over time, and higher for individuals who know Catalan than for individuals who do not know it. Returns to language in terms of reducing over-education may decrease as more individuals know Catalan and job mismatch declines over time. Interestingly, the difference in these returns by actual language knowledge conforms to the theory of comparative advantage, that is, individuals with higher returns to language knowledge are those who actually know it.

We also check for robustness of these results by trying different specifications. Our main results are summarized in Table 4. This table shows the average discrete effects for standard probit and bivariate probit estimations, with and without activity and occupation dummies, for over-education. These alternative specifications do not alter substantially the findings discussed above. However, accounting for endogeneity of Catalan knowledge tends to increase the absolute value of the average discrete effects

on over-education for individuals who read and speak Catalan, but it decreases these same effects for individuals who do not read and speak it. For writing skills bivariate probit estimations yield generally smaller average discrete effects than standard probit estimations, except for women.

In short, selection does matter; neglecting language selection leads to underestimating the effect of Catalan on over-education for reading and speaking skills, and mostly to exaggerating it for writing skills. In no case does this accounting change our result that returns to language in terms of reducing over-education are on average negative but systematically non-significant.

# 6 Conclusions

Catalonia's multilingual economy provides a unique opportunity to assess the incidence of language skills on over-education. Catalan language, formerly confined to informal uses, became co-official in coexistence with Castilian (Spanish) and the language of instruction in the early eighties. This change resulted in an important increase in the knowledge and use of Catalan that did not, however, undermine the intensive use of Castilian in most spheres of communication.

In this study, we inquire whether individuals with deficient language skills may compensate for this disadvantage by accepting jobs for which they have excessive formal skills. At first glance, descriptive evidence suggests that individuals with better language knowledge appear more likely to be over-educated. In other words, individuals who do not know Catalan are more likely to acquire less education. This naive analysis of the data appears to support the 'discouragement hypothesis,' a common presumption in Catalonia and, especially, in the rest of Spain: individuals that are less fluent in the language of instruction are discouraged to proceed with formal education, drop out of school, and at the end, acquire less education.

A more detailed analysis consists on the estimation of a model that controls for

several socio-demographic workers' attributes. In such a model, we find that language knowledge has in fact a negative effect on over-education, which supports the hypothesis of language skills substitution: individuals who are less fluent in Catalan compensate for this disadvantage by acquiring more formal education. This effect is robust to accounting for endogeneity of language knowledge and significant for a representative individual. However, this negative effect is so heterogeneous across individuals that once we compute average discrete effects of language knowledge on over-education, it becomes mostly non-significant.

Our results have implications for contemporary language policy issues, especially in countries that have undergone a change in the language of instruction, such as Morocco from French to Arabic, Wales from English to Welsh, Estonia from Russian to Estonian, or Puerto Rico from English to Spanish. In these economies, individuals who are not fluent in the current language of instruction may get discouraged and acquire less formal education or, on the contrary, compensate for their disadvantage by acquiring more formal education. The two effects seem plausible and could be present in public policy discussions. Our findings based on evidence for Catalonia show that the substitution effect between formal and language skills is weakly predominant over the discouragement effect.

# Appendix

## A.1. Sample Selection

The following table illustrates the importance of the selection criteria in constructing the sample.

| | 1991 | 1996 |
|---|---|---|
| Total sample | 250 000 | 250 000 |
| Only main household members: parents and children | -17 654 | -17 903 |
| Only individuals between 16 and 60 years old | -82 297 | -81 770 |
| Only Spaniards | -5 740 | -4 745 |
| Only if arrival in Catalonia available | | -3 788 |
| Only individuals in the labor force | -47 421 | -44 809 |
| Only if Catalan language variable available | -25 | |
| Selected sample | 96 863 | 96 985 |

## A.2. Definition of the Variables

The explanation on the construction of each variable is presented below.

Over-education.- A worker is defined as over-educated if his/her years of schooling are above the mean educational level of the corresponding occupation plus one standard deviation. Adequately educated workers are those whose educational level is higher than the mean educational level of the corresponding occupation minus one standard deviation and lower than the mean educational level plus one standard deviation. And finally, a worker is under-educated if his/her educational attainment is below the mean education of the corresponding occupation minus one standard deviation.[10]

Schooling.- The census reports the maximum level of studies attained by the individual. To each level, we assign the number of years of schooling.

Age.- It is the census year, 1991 or 1996, respectively, minus the year of birth.

Normalization.- If the individual was younger than 12 years old in 1984, this dummy variable takes the value of one and zero otherwise.

Married.- This variable takes the value of one, if the respondent reports to be currently married; it is zero if the respondent reports to be a widow(er), separated, or divorced.

[10] Among the several ways of measuring fit between workers' acquired and required schooling, two main approaches, subjective and objective, can be distinguished. Subjective definitions are based on individual workers' self-reports on their level of skill utilization. Objective definitions can be classified into two types. In the first type, over-education is assessed by comparing years of education with the average educational level in the worker's current occupation. This method consists of using the distribution of levels of education to construct an over-education index. The second type of objective definition of over-education is based on a comparison between the actual educational level and the job-level requirements. In this paper we use the first type of objective definition.

Residence variables.- The census reports the *municipi* and the Province of residence for each individual. With this information we construct dummies for Lleida, Girona and Tarragona.

Origin variables.-The census reports the *municipi* and the province of birth for each individual. With this information we construct dummies for people who are not born in Catalonia: Andalusia, Valencia, Balearics and La Franja.

YSM (Years since Migration).- The census reports the year of arrival to Catalonia. YSM is the census year minus this number. We also construct the dummy indicating if somebody arrived when s/he was no more than 9 years old.

Municipal variables.- We use the residence variable to assign to each individual the corresponding information of the *municipi*.

Occupational dummies.- These are (1) agricultural occupations (reference group), (2) industrial occupations, (3) trade, service and professional occupations.

Activity dummies.- These are: (1) agricultural activities (reference group), (2) industrial activities, (3) trade activities, (4) service activities.

# References

Acemoglu, D. (2002), 'Technical Change, Inequality, and the Labour Market', *Journal of Economic Literature* **40**, 7–72.

Alba-Ramirez, A. (1993), 'Mismatch in the Spanish Labour Market. Overeducation?', *Journal of Human Resources* **28**, 259–278.

Angrist, J., Chin, A. and Godoy, R. (2006), Is Spanish-only schooling responsible for the Puerto Rican language gap? NBER WP. 12005.
**URL:** *http://www.nber.org/papers/w12005*

Angrist, J. D. and Lavy, V. (1997), 'The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco', *Journal of Labor Economics* **15**, 48–76.

Barrett, A., Bergin, A. and Duffy, D. (2005), The Labour Market Characteristics and Labour Market Impacts of Immigrants in Ireland. IZA DP 1553.

Battu, W. and Sloane, P. J. (2002), Overeducation and Ethnic Minorities in Britain. IZA DP 650.

Becker, G. (1957), *The Economics of Discrimination*, Chicago University Press, Chicago.

Borghans, L. and de Grip, A. (2000), *The Overeducated Worker? The Economics of Skill Utilization*, Borghans, L., and de Grip, A.(eds).

Borjas, G. (1999), The Economic Analysis of Immigration, *in* O. Ashenfelter and D. Card, eds, 'Hanbook of Labor Economics, Volume 3A, Chapter 28', North Holland, Amsterdam, pp. 1697–1760.

Bratsberg, B. and Ragan, J. F. (2002), 'The Impact of Host-country Schooling on Earnings. A Study of Male Immigrants in the United States', *Journal of Human Resources* **37**.

Büchel, F. and Mertens, A. (2004), 'Overeducation, Undereducation and the Theory of Career Mobility', *Applied Economics* **36**.

Casesnoves Ferrer, R., Turell, M. T. and Sankoff, D. (2004), La base démolinguistique pour évaluer l'aménagement linguistique dans un contexte bilingue. X Congrès Linguapax. Diversitat lingüística, sostenibilitat i pau. Fòrum Diàlegs, Barcelona.

Charette, M. and Meng, R. (1994), 'Explaining language proficiency. Objective versus self-assessed measures of literacy', *Economic Letters* **44**, 313–321.

Chiswick, B. and Miller, P. (2002), 'Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle', *Journal of Political Economy* **15**, 31–57.

Dekker, R, d. A. and Heijke, H. (2002), 'The Effects of Training and Overeducation on Career Mobility in a Sequential Labour Market', *International Journal of Manpower* **23**.

Dolton, P. and Vignoles, A. (2000), 'The Incedence and Effects of Overeducation in the Uk Graduate Labour Market', *Economics of Education Review* **19**, 179–198.

Dupuy, A. and de Grip, A. (2002), Do Large Firms Have More Opportunities to Substitute between Skill Categories than Small Firms? Aarhus Centre for Labour Market and Social Research, working paper 02-01.

Freeman, R. (1976), *The Overeducated American*, Academic Press, New York.

Gordon, R. J. (2005), *Ethnologue: Languages of the World, Fifteenth edition*, SIL International. Online version: http://www.ethnologue.com/, Dallas, Tex.

Green, C., Kler, P. and Leeves, G. (2004), Overeducation and Assimilation of Recently Arrived Immigrants: Evidence from Australia. University of Queensland, WP 4.

Grin, F. (2003), 'Language planning and economics', *Current Issues in Language Planning* **4**(1), 1–66.

Grin, F. and Vaillancourt, F. (1998), Language Revitalisation Policy: An Analytical Survey. Treasury Working Paper 98/6. Mimeo.

Groot, W. (1993), 'Overeducation and the Returns to Enterprise-related Training', *European Economic Review* **12**, 299–309.

Groot, W. (1996), 'The Incidence of, and Returns to Overeducation in the UK', *Applied Economics* **28**, 1345–1350.

Groot, W. and Maassen van den Brink, H. (1997), 'Allocation and the Returns to Overeducation in the United Kingdom', *Education Economics* **5**, 169–183.

Hartog, J. (2000), 'Overeducation and Earnings: Where Are We, Where Should We Go?', *Economics of Education Review* **19**, 131–147.

Jovanovic, B. (1979), 'Job Matching and the Theory of Turnover', *Journal of Political Economy* **87**, 972–990.

Linsley, I. (2005), Overeducation in the Australian Labour Market: Its Incidence and Effects. The University of Melbourne, WP 939.

Maddala, G. S. (1983), *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.

Niesing, W., van Praag, B. and Veenman, J. (1994), 'The Unemployment of Ethnic Minority Groups in the Netherlands', *Journal of Econometrics* **61**.

Oosterbeek, H. (2000), 'Introduction to Special Issue on Over-schooling', *Economics of Education Review* **19**.

Raynauld, A. and Marion, P. (1972), 'Une analyse économique de la disparité inter-ethnique des revenus', *Revue Économique* (23), 1–19.

Rendon, S. (2006), The Catalan Premium: Work and Language in Catalonia. Forthcoming in the *Journal of Population Economics*.

Robst, J. (1995), 'Career Mobility, Job Match, and Overeducation', *Eastern Economic Journal* .

Rosen, S. (1972), 'Learning and Experience in the Labour Market', *Journal of Human Resources* **7**, 326–342.

Sabourin, C. and Bernier, J. (2001), Government responses to language issues. international examples. Office of the Languages Commissioner of Nunavut. Mimeo.

Schaafsma, J. and Sweetman, A. (2001), 'Immigrant Earnings: Age at Immigration Matters', *Canadian Journal of Economics* **34**.

Sicherman, N. (1991), 'Overeducation in the Labor Market', *Journal of Labor Economics* **9**, 101–122.

Sicherman, N. and Galor, O. (1990), 'A Theory of Career Mobility', *Journal of Political Economy* **98**, 160–192.

Spence, M. (1973), 'Job Market Signalling', *Quarterly Journal of Economics* **87**, 355–374.

Thurow, L. (1975), *Generating Inequality: Mechanisms of Distribution in the U.S Economy*, New York: Basic Books, New York.

Vila, F. X. and Vial, S. (2001), Escola i ús. les pràctiques lingüístiques de lŠalumnat de 2n nivell de cicle superior dŠeducació primària de catalunya en situacions qusi-espontànies. Mimeo.

Wieling, M. and Borghans, L. (2001), 'Discrepancies between supply and demand and adjustment processes in the labour market', *Labour* **15**, 33–56.

Woolard, K. A. (1989), *Double Talk. Bilingualism and the politics of ethnicity in Catalonia*, Stanford University Press cop., Stanford, California.

Table 1: Summary Statistics

| Census Year | 1991 | | 1996 | |
|---|---|---|---|---|
| Gender | Men | Wom. | Men | Wom. |
| Education | | | | |
| %Under-educated | 14.1 | 12.5 | 9.1 | 8.4 |
| %Adequately Educated | 71.5 | 75.9 | 81.2 | 80.7 |
| %Over-educated | 14.4 | 11.6 | 9.7 | 10.9 |
| Average Years of Schooling of the | | | | |
| Under-educated | 3.9 | 4.6 | 5.9 | 5.7 |
| Adequately Educated | 7.7 | 8.6 | 8.5 | 9.3 |
| Over-educated | 12.0 | 12.8 | 12.8 | 13.6 |
| Average Age of the | | | | |
| Under-educated | 46.1 | 43.4 | 45.2 | 42.4 |
| Adequately Educated | 37.8 | 34.4 | 38.4 | 36.2 |
| Over-educated | 34.3 | 31.5 | 36.3 | 33.5 |
| % Read & Speak Catalan among the | | | | |
| Under-educated | 39.8 | 50.3 | 57.1 | 64.7 |
| Adequately Educated | 64.9 | 75.2 | 75.0 | 83.8 |
| Over-educated | 77.1 | 82.1 | 88.9 | 92.8 |
| % Write Catalan among the | | | | |
| Under-educated | 17.2 | 26.8 | 28.8 | 39.1 |
| Adequately Educated | 37.1 | 51.5 | 46.0 | 60.7 |
| Over-educated | 54.0 | 65.2 | 67.6 | 80.6 |
| % Over-education among those who | | | | |
| Read & Speak Catalan | 17.6 | 13.0 | 11.5 | 12.1 |
| Do not Read & Speak Catalan | 8.9 | 7.7 | 4.2 | 4.6 |
| Write Catalan | 21.2 | 15.1 | 14.0 | 14.3 |
| Do not Write Catalan | 10.5 | 8.1 | 5.9 | 5.4 |
| %Over-educated: Total | 14.4 | 11.6 | 9.7 | 10.9 |
| %Over-educated: Service Sector | 17.7 | 12.8 | 11.6 | 12.0 |
| % Women | | 34.5 | | 37.3 |
| % Women working in the Service Sector | | 56.6 | | 57.0 |
| % Married | 69.6 | 59.5 | 67.0 | 59.9 |
| % Normalized | 3.1 | 3.9 | 10.2 | 12.7 |

Table 1 (cont): Summary Statistics

| Census Year | 1991 | | 1996 | |
|---|---|---|---|---|
| Gender | Men | Wom. | Men | Wom. |
| Residence | | | | |
| % Barcelona | 86.3 | 77.7 | 75.0 | 75.8 |
| % Girona | 8.7 | 8.9 | 9.1 | 9.7 |
| % Lleida | 5.9 | 5.5 | 6.3 | 5.7 |
| % Tarragona | 9.1 | 7.9 | 9.6 | 8.8 |
| Municipi | | | | |
| % writes Catalan in Municipi | 39.8 | 40.5 | 46.0 | 46.4 |
| % Catalan-born in Municipi | 67.3 | 67.7 | 68.5 | 68.6 |
| % work in Services in Municipi | 51.6 | 52.6 | 57.2 | 58.2 |
| % employed over population in Municipi | 37.2 | 37.5 | 36.4 | 36.4 |
| Origin | | | | |
| % not born in Catalonia | 37.6 | 30.2 | 31.0 | 24.9 |
| % born in Andalusia | 24.1 | 18.3 | 19.5 | 14.4 |
| % born in Valencia-Balearics | 1.2 | 1.2 | 1.1 | 1.1 |
| % born in Franja | 0.5 | 0.5 | 0.4 | 0.4 |
| % arrived age $\leq 9$ | 11.0 | 11.1 | 10.0 | 9.6 |
| Years since migration if not born in Cat | 25.0 | 23.6 | 28.4 | 26.8 |

Table 2. Over and Under–Education Language Discrete Effects and Premia by Reading and Speaking, and Writing Skills (Standard errors in small fonts)

| Catalan Skill | Reading and Speaking | | | | Writing | | | |
|---|---|---|---|---|---|---|---|---|
| Census Year | 1991 | | 1996 | | 1991 | | 1996 | |
| Gender | Men | Wom. | Men | Wom. | Men | Wom. | Men | Wom. |
| OVER-EDUCATION | | | | | | | | |
| Basic | | | | | | | | |
| Representative | -2.14 | -4.88 | -0.15 | -2.06 | -1.78 | -3.63 | -0.09 | -1.06 |
| | 0.21 | 0.45 | 0.03 | 0.35 | 0.15 | 0.29 | 0.02 | 0.16 |
| Average | -3.54 | -4.44 | -2.75 | -5.67 | -3.17 | -3.52 | -1.89 | -3.48 |
| | 3.51 | 5.20 | 3.52 | 6.52 | 3.18 | 4.31 | 2.45 | 4.11 |
| Occupation and Activity dummies | | | | | | | | |
| Representative | -0.30 | -0.65 | 0.00 | -0.29 | -0.25 | -0.52 | 0.00 | -0.15 |
| | 0.07 | 0.15 | 0.00 | 0.09 | 0.05 | 0.10 | 0.00 | 0.04 |
| Average | -1.33 | -2.02 | -1.04 | -3.16 | -1.17 | -1.73 | -0.72 | -1.88 |
| | 1.61 | 2.72 | 1.46 | 3.69 | 1.40 | 2.35 | 1.01 | 2.22 |

Table 3: Probability of Over–Education and Average Discrete Effect (in %) by Catalan skills. Standard errors in small fonts

| Census Year | 1991 | | | | 1996 | | | |
|---|---|---|---|---|---|---|---|---|
| Gender | Men | | Women | | Men | | Women | |
| Knowledge | Knw. | D.knw | Knw. | D.knw | Knw. | D.knw | Knw. | D.knw |
| READING AND SPEAKING | | | | | | | | |
| Actual | 17.62 | 8.93 | 13.03 | 7.66 | 11.50 | 4.23 | 12.11 | 4.63 |
| Pred: Know | 19.21 | 8.38 | 15.06 | 5.65 | 13.83 | 4.18 | 17.03 | 4.77 |
| | 25.28 | 17.40 | 22.93 | 14.34 | 22.96 | 13.69 | 24.61 | 14.37 |
| Pred: DKnow | 21.06 | 9.22 | 18.06 | 6.83 | 15.31 | 4.57 | 20.79 | 5.70 |
| | 26.60 | 18.50 | 25.22 | 16.02 | 24.19 | 14.44 | 27.16 | 15.92 |
| **Av. D. Effect** | -1.85 | -0.84 | -3.00 | -1.18 | -1.48 | -0.39 | -3.76 | -0.94 |
| | 1.84 | 1.36 | 3.31 | 2.14 | 1.86 | 0.97 | 4.07 | 2.06 |
| WRITING | | | | | | | | |
| Actual | 21.19 | 10.47 | 15.09 | 8.06 | 14.03 | 5.86 | 14.32 | 5.41 |
| Pred: Know | 23.35 | 10.09 | 18.63 | 6.68 | 17.08 | 6.12 | 20.19 | 6.49 |
| | 26.63 | 19.16 | 24.68 | 15.63 | 24.74 | 16.08 | 25.75 | 16.52 |
| Pred: DKnow | 24.86 | 10.87 | 21.45 | 7.92 | 18.06 | 6.49 | 22.76 | 7.32 |
| | 27.47 | 20.04 | 26.41 | 17.18 | 25.40 | 16.64 | 27.17 | 17.65 |
| **Av. D. Effect** | -1.50 | -0.78 | -2.82 | -1.24 | -0.99 | -0.37 | -2.57 | -0.83 |
| | 1.24 | 1.13 | 2.63 | 2.03 | 1.04 | 0.74 | 2.42 | 1.52 |

Table 4: Over and Under–Education Average Discrete Effect (in %) by Catalan skills Standard errors in small fonts

| Census Year | 1991 | | | | 1996 | | | |
|---|---|---|---|---|---|---|---|---|
| Gender | Men | | Women | | Men | | Women | |
| Knowledge | Knw. | D.knw | Knw. | D.knw | Knw. | D.knw | Knw. | D.knw |
| OVER-EDUCATION | | | | | | | | |
| READING AND SPEAKING | | | | | | | | |
| PROBIT | -4.45 | -2.05 | -5.89 | -2.19 | -3.38 | -0.99 | -6.57 | -1.68 |
| | 3.27 | 2.75 | 5.01 | 3.24 | 3.62 | 2.30 | 6.25 | 3.80 |
| BIPROBIT | -5.16 | -2.02 | -6.03 | -1.78 | -4.08 | -0.92 | -7.65 | -1.44 |
| | 4.02 | 2.80 | 5.58 | 2.84 | 4.60 | 2.23 | 7.88 | 3.53 |
| PROBIT+OA DUMMIES | -1.63 | -0.85 | -2.55 | -1.20 | -1.27 | -0.42 | -3.61 | -1.20 |
| | 1.56 | 1.36 | 2.70 | 2.15 | 1.53 | 1.03 | 3.67 | 2.51 |
| BIPROBIT+OA DUMMIES | -1.85 | -0.84 | -3.00 | -1.18 | -1.48 | -0.39 | -3.76 | -0.94 |
| | 1.84 | 1.36 | 3.31 | 2.14 | 1.86 | 0.97 | 4.07 | 2.06 |
| WRITING | | | | | | | | |
| PROBIT | -4.81 | -2.22 | -6.10 | -2.04 | -2.87 | -1.05 | -4.89 | -1.45 |
| | 2.88 | 2.69 | 4.34 | 2.94 | 2.59 | 1.94 | 4.03 | 2.79 |
| BIPROBIT | -4.72 | -2.17 | -5.92 | -1.88 | -2.87 | -0.96 | -5.25 | -1.33 |
| | 2.88 | 2.58 | 4.46 | 2.65 | 2.71 | 1.79 | 4.70 | 2.64 |
| PROBIT+OA DUMMIES | -1.68 | -0.87 | -2.66 | -1.20 | -1.05 | -0.42 | -2.53 | -0.94 |
| | 1.36 | 1.25 | 2.41 | 1.96 | 1.08 | 0.84 | 2.24 | 1.71 |
| BIPROBIT+OA DUMMIES | -1.50 | -0.78 | -2.82 | -1.24 | -0.99 | -0.37 | -2.57 | -0.83 |
| | 1.24 | 1.13 | 2.63 | 2.03 | 1.04 | 0.74 | 2.42 | 1.52 |

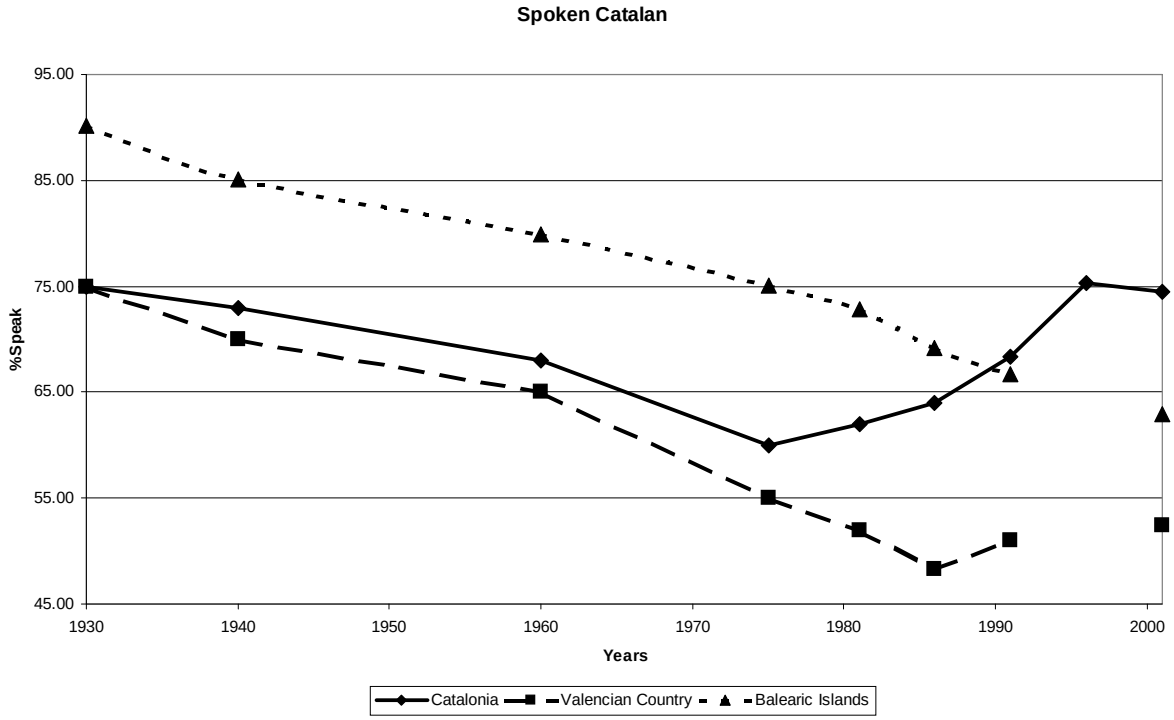


Figure 1: Spoken Catalan in Catalonia, Valencia and Balearics

(Source: Vallverdú 1990 and Census data)