Returns to Foreign Languages of Native Workers in the EU[*]

CORE Discussion Paper 2007/21

Victor Ginsburgh
ECARES, Université Libre de Bruxelles and
CORE, Université catholique de Louvain

Juan Prieto-Rodriguez
Department of Economics
University of Oviedo

December 2006

Abstract

Most papers on returns to languages are concerned with immigrants. We use the European Community Household Panel Survey (ECHP) to infer returns on non-native languages by non-immigrants in nine countries of the European Union. We differ from the few other studies that deal with the same problem in three respects. First, we correct for time-dependent measurement errors in self-reporting as suggested by Dustmann and Van Soest and find that the resulting IV estimates are much larger than those obtained by OLS. We also suggest that there is little room for time-persistent errors and heterogeneity, and that therefore our estimates should not suffer from the other usual biases. Secondly, instead of using a dummy for each language, we use the ratio of the population that is not proficient in a language in each country considered. Finally, we estimate instrumental variable quantile regressions to illustrate how returns to languages vary at different points of the distribution of earnings.

[*] Ginsburgh is also member of ECORE, the newly created association between CORE and ECARES. We are grateful to David Veredas for many enlightening discussions, and to participants of the ECORE Conference Challenges of Multilingual Societies, Brussels June 9-10, 2006.

1. **Introduction and Background**

Globalization, increasing international trade flows, the establishment of new free-trade areas and the efficient performance of international "common markets" result from improvements in transportation, but also and probably even more so, from the ability of economic agents to communicate. Given new communication technologies, the only restrictions left are due to linguistic barriers. Although English has become the current *lingua franca* (as was, in other times, Latin, and somewhat later, French) and is understood by 1.5 to 1.8 billion citizens (Crystal, 2001), countries whose official tongues are not English represent a large share of the world's GDP and their inhabitants may substantially improve their human capital if they know languages other than their own mother tongue (or English). However, one may expect that the more the official language of a country is international, the less important the economic advantage of its inhabitants to learn other languages, and the smaller the domestic language in a country, the higher the number of foreign languages speakers.[1] Obviously, not all languages will be equally rewarded by the labor market due both to supply and demand factors.

Though languages are important in international transactions, the domestic relevance of language skills -- for example, the importance for immigrants to know the language spoken in the country to which they move -- is the topic of many papers. Not surprisingly, therefore, most papers on the returns of language proficiency deal with the skills of immigrants in "traditional" immigration countries such as Australia, Canada, Germany, Israel, the United Kingdom and the United States.[2] Only a couple of papers consider the case of natives in multilingual societies, such as Canada (Shapiro and Stelcner, 1997), Hungary (Galasi, 2003), Luxemburg (Klein, 2003), Switzerland (Cattaneo and Winkelman, 2003) and the United States (Fry and Lowell, 2003). Countries of the European Union before its last enlargement to the East (EU 15) are the subject of a paper by Williams (2005).

A second strand of research is concerned with the determinants of language acquisition by immigrants (Chiswick and Miller, 1995, 1996, 2003, Dustmann, 1994, Dustmann and Van Soest, 2001). Language proficiency is assumed to depend on three sets of factors: economic incentives linked to expected returns, individual efficiency in language acquisition, and exposure to the language prior to (and after) immigration.

We concentrate on the returns of multilingualism in a certain number of countries of the European Union, thus following on three papers that considered the same issue. Galasi (2003) uses two Hungarian surveys of young career beginners who graduated from public higher education as full-time students in 1998 and 1999. His paper is devoted to the returns to

[1] See e.g. Ginsburgh, Ortuno-Ortin and Weber (2005) for a model and some empirical estimates.

[2] Australia (Chiswick and Miller, 1995), Canada (Abbott and Beach, 1992, Aydemir and Skuterud, 2005, Chiswick and Miller, 1995), Germany (Dustmann and Van Soest, 2002), Israel (Beenstock, Chiswick and Repetto, 2001, Berman, Lang and Siniver, 2003, Chiswick and Miller, 1995, Chiswick, 1998), the United Kingdom (Leslie and Lindley, 2001), and the United States (Bleakey and Chin, 2004, Bratsberg, Ragan and Nasir, 2002, Chiswick and Miller, 1995, 2002, Hellerstein and Neumark, 2003).

education rather than to languages. Due to this, he merely corrects for the usual bias in estimating the returns to education equation but not for the similar bias with respect to language proficiency. He finds that speaking English or German increases wages by some 6 and 4 percent, respectively, although the results are not very robust to alternative specifications (in some cases, he even obtains negative, though insignificant, returns).

In his paper on the returns to languages in Luxemburg, Klein (2003) concludes that there is little advantage to be proficient in any of the official languages (French, German and Luxemburgish), but it pays to know English, the only truly "foreign" language.

The closest paper to ours is Williams (2005) who uses data from the European Community Household Panel Survey (ECHP) between 1994 and 1999. The information results from answers to the following question included in the surveys: "Does your work involve the use of a language other than (the official language of the country)?" Williams runs ordinary least squares Mincer-type regressions where the explanatory variables include a broad array of socio-economic indicators. To capture the effect of language knowledge, a dummy is introduced for each of the following languages if it is cited as the first foreign language used at the workplace of the respondent: English, French, German, Spanish, Italian, Dutch, “all other”. In an alternative specification, all languages are pooled, and only one coefficient is estimated. Equations are run for each EU 15 member state separately (with the exception of Sweden) in 1996. The results indicate that in Austria, Finland, Italy, Spain and the Netherlands, English is the only language that yields a significant return. However, substantial returns are also found for French in Denmark, Luxemburg, Greece and Portugal, while German generates positive and significant returns in Belgium, Luxemburg and France, Spanish does so in France, Italian in Luxemburg and Portuguese, and Dutch in Belgium. In the United Kingdom, no second language is rewarded. Languages add 5 to 20 percent to earnings, depending on the country and the language considered.

Our paper estimates returns to languages in nine European countries in 2001, a more recent year of the same panel survey as the one used by Williams. We differ in three respects from Williams. First, we try to take into account the endogeneity issue between languages and earnings. Secondly, instead of representing proficiency in languages by dummy variables (or by the level of efficiency), we use an indicator which makes a link with the “usefulness” of the language or the frequency with which it is spoken in a given country that is, the share of the population which does *not* know the language (the *disenfranchisement* rate). For individuals who only use the language of the country (French in France, for example), this share is zero (or close to 0 if the country hosts immigrants who do not speak the official language), while for a language that nobody knows but is useful at the workplace, it will be equal to 1. The parameter associated with the variable allows retrieving returns to languages for which one knows the disenfranchisement rate, and as long as the estimated parameter is positive, the returns will be larger for languages that are less common. Thirdly, following Chernozhukov and Hansen (2004, 2005 and 2006) instrumental variable quantile regression estimator, we also estimate instrumental variable quantile regressions which illustrate how returns to languages vary at different points of the distribution of earnings.

The paper is organized as follows. In Section 2, we discuss the data used, the model, and the econometric problems that have to be faced in estimating such a model, as well as the results of the estimation of "mean" returns obtained by usual (instrumental variables) techniques. In Section 3, we turn to the results obtained by quantile (instrumental variables) regressions which give more detailed information on the relation between language proficiency and earnings. Section 4 concludes the paper.

2. Data, the Model and Econometric Issues

Data

The database that we use is the European Community Household Panel (ECHP), which contains information on a panel of individuals in 15 European countries from 1994 to 2001. This information is homogenous across countries since the surveys were coordinated by EUROSTAT, although the sample sizes vary across countries and years. The surveys describe the socio-economic characteristics of individuals older than 16, grouped by households, including personal characteristics, family structure, current employment, education and training, labor status, wages, family income from sources other than wages and salaries, region of residence, and languages used at the workplace.

Between 1994 and 1999, questions about language skills concerned first, second and third foreign languages used at the workplace. These questions were dropped in Belgium, Germany, Luxemburg and the United Kingdom between 1997 and 1999. Sweden joined the survey in 1997 but did not include the questions on languages. In 2000 and 2001, questions were formulated somewhat differently, and moved to another part of the questionnaire. Workers were asked to report on the main and second languages used "at their main job."[3] We were faced with the following issues and possibilities: (a) we found inappropriate to merge both parts of the panel, i.e. 1994-1999 and 2000-2001, since the essential questions on languages changed after 2000; (b) we could have used panel techniques on the 1994-1999 waves, with the exception of Belgium, Germany, Luxemburg and the United Kingdom for which only three years were available; (c) the waves 2000-2001 are more recent, the information is consistent, but running panel techniques with only two years is not very useful. We preferred to use the information contained in the 2000 survey to control for non-persistent time components of the error term.[4]

[3] In Spain, the question was formulated "main and second *foreign* language." These questions were not part of the British and Swedish surveys and are not available for the Netherlands in 2001.

[4] Since we also run quantile regressions, this also allows us to compare the results of IV and quantile estimation.

Though both parts of the panel include 15 countries, the questions on languages failed to be included for the Netherlands, Sweden and the United Kingdom. We left out Luxemburg and Belgium given their multilingual situation. Ireland was also dropped since the survey contains a very small number of respondents (less than 2 percent) who need a foreign language at their workplace.[5]

We also excluded immigrants from the sample, since the mechanisms that govern language acquisition differ for non-natives.[6] Table 1 contains information on the number of observations in each country, as well as on the number of individuals who report needing one or several among the main European languages, that is English, French, German, Italian, Spanish and Dutch, as well as the official language of the country.

Specification of the returns to languages equation

The objective is to estimate the effect of language knowledge (and use at the workplace) on earnings. The standard (language-augmented) Mincer-type equation can be specified as:

$$\text{(1)} \qquad \ln w_i = x_i\beta + D_i\gamma + u_i$$

for individuals $i = 1, 2, ..., N$, where the vector β and the scalar γ are parameters, w_i represents the wage rate, x_i is a vector of exogenous variables and u_i is a random error. In the case of immigrants, D_i is usually a dummy variable that takes the value 1 if individual i reports being proficient in the language of the country for which the equation is estimated, and 0 otherwise. In some cases, the variable represents various degrees of self-assessed proficiency in the language,[7] but even then the variable is often coded as a dummy, equal to 1 if proficiency is larger than a given threshold, and 0 otherwise.[8] When the equation is run to assess the returns of a foreign language, the dummy usually represents self-assessed knowledge of the non-native language that individual i knows. In case of several languages, each language is represented by a dummy (Galasi, 2003, Williams, 2005). Bratsberg, Ragan and Nasir (2002) introduce a variable that takes the value 1 once an individual acquires the American citizenship.

There is an important difference between our model and the usual one, in which the “language” variable is concerned with language proficiency of an individual. In our case, individuals self-report the language(s) used at their main workplace. Therefore, once a

[5] We did run regressions for Ireland, but the standard errors of the estimated parameters were very large, and almost no coefficient of interest was significantly different from zero.

[6] Women were also excluded since this would have increased heterogeneity and it is out of the scope of this paper to proceed to gender comparisons.

[7] See Berman, Lang and Siniver (2003), Chiswick and Miller (2002), Fry and Lowell (2003),

[8] See e.g. Dustmann and Van Soest (2002).

language is mentioned, it measures not only knowledge, but also what the firm needs, that is, the demand side. This will have an influence on the estimation procedure.

Econometric issues

The returns to languages equation is subject to unobserved heterogeneity similar to the one faced in the returns to education literature, which tries to assess the effect of education on earnings: both education and earnings may be dependent on unobservable individual skills and talent. Explanatory variables, in particular D_i are correlated with the error term in equation (1). This yields biased ordinary least squares estimates, and overestimates the parameters of interest. The solution pursued in the literature on private returns to education is to use instruments for education (such as time of admission, see Angrist and Krueger, 1991), or to rely on natural experiments (such as educational outcomes and earnings of identical twins, see Ashenfelter and Krueger, 1994).

The same issue is pervasive in the returns to language acquisition equation, and needs instrumental variable estimation. Chiswick and Miller (1995) use the following instruments for language fluency: number and age of children, individual married overseas, and minority-language concentration of the place of residence, defined as the proportion of the population in the region who report the same minority language, and in which the respondent lives.[9] The latter variable has been criticized since it may be considered a choice variable, and as observed by Bleakey and Chin (2004), regional characteristics correlated with the concentration ratio may have effects on earnings. Dustmann and Van Soest (2002) use the father's education level. Bleakey and Chin (2004) use age at arrival of the immigrant interacted with a dummy for non-English speaking country of origin as identifying instrument. This allows taking into account the possible effect of easier adaptation of an immigrant when he moves at a young age. Galasi (2003) who studies the return of foreign languages in Hungary uses scores obtained by individuals when they took their higher-education admission exam.

A second issue, related to misclassified language indicators, is discussed by Dustmann and Van Soest (2001).[10] In panel (or cross-section) data, language ability is usually self-reported. This is certainly so when individuals report on their level of language proficiency using an ordinal scale, but also when their answer is dichotomous (yes or no). Two types of errors affect these variables: a purely random error, that is independent of time and an error

[9] Chiswick and Miller also use the language of the country in which the immigrant was born.

[10] This is a problem that is not raised in the economics of education literature, probably because individuals remember (or do not cheat about) their education level.

that is time-dependent, since an individual may have the same tendency to over- or under-report.[11]

To make this clear, we follow Dustmann and Van Soest, and rewrite (1) as:

$$(2) \qquad \ln w_{it} = x_{it}\beta + D_{it}\gamma + \alpha_i + v_{it}$$

where time subscripts as well as unobserved individual-specific effects α_i are introduced, which may be correlated with D_{it}, and where the v_{it} are random error terms that are uncorrelated with x_{it} and D_{it}. Since D_{it} is self-reported, it is usually measured with errors. Let the true value D_{it} and the observed value $\tilde{D}_{it}$ be related as follows:

$$(3) \qquad \tilde{D}_{it} = D_{it} + \eta_{it} + \xi_i$$

where η_{it} represents a time-independent unsystematic error and ξ_i represents a time-persistent error. Substitution of (3) in (2) leads to (4) and clarifies the possible correlations between the error terms and the language variable:

$$(4) \qquad \ln w_{it} = x_{it}\beta + (\tilde{D}_{it} - \eta_{it} - \xi_i)\gamma + \alpha_i + v_{it} = x_{it}\beta + \tilde{D}_{it}\gamma + (\alpha_i - \gamma\xi_i) + (v_{it} - \gamma\eta_{it})$$

Estimating (4) by ordinary least squares yields biased results of γ if the error term $[(\alpha_i - \gamma\xi_i) + (v_{it} - \gamma\eta_{it})]$ is correlated with $\tilde{D}_{it}$. This can be due to measurement errors (in which case η_{it} and ξ_i are correlated with $\tilde{D}_{it}$) or to correlation between the unobserved individual-specific α_i and $\tilde{D}_{it}$.

To deal with the time-dependent measurement error, Dustmann and Van Soest suggest using leads and lags of self-reported language fluency as instruments for current fluency. They note, however, that this does not eliminate time-persistent errors, which introduce heterogeneity, and can be taken care of by the introduction of additional control variables (household characteristics, such as partner variables if one believes that mating is assortative). Finally, in order to remove all types of correlations simultaneously, they suggest using as instrument the education level of parents. The results obtained by Dustmann and Van Soest show that the bias due to time-independent misclassification errors is much larger than the one due to the usual type of unobserved heterogeneity. They also find that the bias due to time-persistent errors is quite small. This would imply that endogeneity problems generated by the first component of the error term in (4) could be considered negligible.

[11] Dustmann and Van Soest (2001) formulate a model that allows isolating the two types of errors if panel data are used to estimate the returns equation.

Specification used in this paper

In the equations that we estimate, we wanted to keep the model as close as possible to the Mincerian model used in the literature on returns on education. The dependent variable is the natural logarithm of the wage rate. The vector x_i contains the following control variables: two dummy variables that represent higher education (university) and secondary level education, respectively; the number of years of job tenure and its square; the number of years of potential experience and its square.

To introduce specific languages effects we use a unique scalar variable D_i which measures the country-wide "disenfranchisement rate" (Ginsburgh and Weber, 2005) of the language that individual i reports to be using at his workplace.[12] This rate represents the percentage of the population that is not proficient in a foreign (or native) language, and is based on the results of a survey commissioned in 2000, by the Directorate of Education and Culture of the EU (See INRA, 2001). In each of the EU15 countries, 1,000 interviews were conducted on language proficiency. The information used by Ginsburgh and Weber takes into account answers to the following two questions:

"(a) What is your mother tongue? (note to the interviewer: do not probe; do not read [the list of languages] out; if bilingual, state both languages);
(b) What other languages do you know? (show card [containing a list of languages]; read out; multiple answers possible)."

There were four possible choices for (b), and the assumption made by Ginsburgh and Weber was that the first two choices that came to the mind of the individual interviewed were the languages that he knew best. There were also questions on whether the knowledge of each tongue mentioned was "very good," "good" or "basic," but the answers were not taken into account.

D_{it} is thus a scalar variable which takes the value 0 for an individual who uses no foreign language at his workplace[13] and the value taken by the disenfranchisement rate for the five most widely internationally spoken languages in the European Union before the 2004 enlargement (English, French, German, Italian and Spanish) in Austria, Denmark, Finland, France, Germany, Greece, Italy, Portugal and Spain for which returns on languages are estimated. The scalar nature of the D_{it} variable will also be useful in our quantile variable estimation procedure, which allows for only one variable that can be instrumented. Disenfranchisement rates for the five languages examined can be found in Table 2.

[12] If the individual reports that he uses several languages, the variable takes the value of the language for which the disenfranchisement rate is the largest.

[13] To avoid problems due to the possible presence of immigrants who do not know the official language the samples used in this research do not include immigrants. Hence, official languages disenfranchisement rates are equal to zero.

The expected sign for parameter γ is positive, since the less a foreign language is known, that is, the higher the disenfranchisement rate, the higher is the expected return to the language. This formulation has the advantage that all foreign languages are subsumed by a unique variable, while the effect of each language can easily be retrieved by multiplying γ by the disenfranchisement rate of the language in a given country.

Recall that D_{it} can vary over time for several reasons: (a) changes within the firm (the firm starts exporting to some new markets or stops doing so between *t-1* and *t*) reported by individual *i*; (b) an individual surveyed in 2000 has moved to another firm, where the language requirements are different; (c) self-reporting errors.

To deal with time-varying errors, we follow Dustmann and Van Soest and instrument D_{it} by its lagged (2000) value. Since in our case D_{it} represents language(s) used at the workplace, and not self-reported linguistic abilities, there is little room for time-persistent measurement errors as, in principle at least, the reported answer can be considered objective. D_{it} measures thus a "demand" side requirement by the firm that hires the worker. Then, even if the variable is correlated with the error term, we can quite safely assume that the correlation is not related to unobserved ability of the individual since supply-side and demand-side random terms can be assumed uncorrelated. Therefore, it is probably rather innocuous to ignore the time-persistent individual effect α_i appearing in the equation.

Table 3 gives an overview of the changes in self-reporting between the two surveys. Details are available for all six languages in each country, but we only report on the total number of switches in percentage of the total number of observations used in the regressions for 2001. As can be seen, differences may be quite significant, accounting for as much as 24.3 and 18.1 percent in Denmark and Finland, respectively. The lower part of Table 4 presents the regression coefficient of D_i on $D_{i,-1}$ (and other exogenous variables) in the regression in which D_i is instrumented (first stage of GMM estimations). The results show that the R-squared of these equations are not close to 1 (except for Germany), which illustrates again that there are quite large differences in reporting language use between the two waves of the survey.

The survey contains unfortunately no variables that could be used as instruments to correct for unobserved heterogeneity,[14] but the results by Dustmann and Van Soest (2002) indicate that correcting for misclassification already produces meaningful results. Moreover, as was pointed out earlier, it is likely that there is no reason to expect such type of

[14] We tried to use partner's education, but this had the effect of reducing the sample size (since we lose workers who have no partner). Moreover, according to Anderson canonical correlations likelihood-ratio test and the Cragg-Donald chi-squared test statistic, models estimated using only partner's education as exogenous instrument are underidentified for all countries with the exceptions of Austria and Spain. In addition, when including both lagged disenfranchisement rate and partner's education as exogenous instruments models are overidentified and, following Hall and Peixe (2003), partner's education is a redundant instrument since the asymptotic efficiency of the estimation is not improved by using this instrument.

.

heterogeneity in our case. Additional insights are given in the Appendix, where some estimations controlling for unobserved ability are presented, confirming our expectations.

Results

GMM estimation results of Equation (4) with t = 2001 and where D_{it} is instrumented by the lagged value $D_{i,t\text{-}1}$ are reported in Table 4. As can be checked the parameter of interest, after instrumentation, is positive for all nine countries, and is estimated with small standard errors, with the exception of Denmark, where it is significantly different from zero at the 7 percent probability level only. Some heterogeneity between countries is observed with the largest values obtained for Finland, France, Germany, Portugal and Spain and the lowest for Denmark. Other values are intermediate. In Table 5, these coefficients are transformed into returns to languages for the five most widely used languages in Europe. Since the dependent variable (wage rate) is defined in logarithms, the parameters can be interpreted as reflecting the percentage increase of the wage rate. Returns will be large if either the estimated country parameter or/and if the disenfranchisement rate for the language is large in the specific country, since returns are obtained as the product of the parameter and the disenfranchisement rate. Note however that in several countries, the number of workers who use a foreign language is quite small, and sometimes there is no such worker in the survey. Our specification allows nevertheless estimating *potential* returns (given that a firm needs an individual who knows the language), but to make this clear, returns for languages for which there are less than ten observations are in italics. Returns are large in Spain, France, Portugal, as well as for Romance languages in Germany, and they are far from being negligible in other countries. Though English is spoken in most countries, it still is in demand in France, Portugal and Spain, but so are French, German, Italian and Spanish in most countries.

Table 6 compares OLS and instrumental variables results. As is the case in other papers (Bleakey and Chin, 2004, Galassi, 2003, and Dustmann and Van Soest, 2001, 2002), IV estimation leads to much larger returns than OLS. In the presence of measurement errors, one can expect OLS to generate a downward bias. On the other hand, OLS produce an upward bias in the presence of unobserved heterogeneity. Our result can be considered as additional evidence for the larger importance of misclassification errors.

As we already mentioned, the number of other variables is small on purpose, in order to remain close to a Mincer-type specification. In fact, usual control variables, such as occupation or firm size, may be (positively) correlated with D_i and their inclusion could lead to underestimate foreign language returns. Both higher and secondary education have positive effects that are all significantly different from zero at conventional probability levels, but the effect of secondary education is smaller. Years of tenure have a positive influence, though it is not significantly different from zero in Austria, Finland, and Portugal. The relation is linear in other countries: the squared term is not significant in most cases, with the exception of

Portugal. Potential experience (number of years after last degree) also has, as expected, a positive impact, but returns are decreasing (the squared term is negative and significant in all countries). These results are close to those obtained by other researchers.

Williams (2005) who is interested, as we are, in the returns to languages in Europe also finds large effects in some cases. He uses languages dummies and estimates returns on log wages that are as large as 0.46 for Spanish in Denmark, 0.32 and 0.44 for French in Greece and Denmark and 0.96 for Italian in Portugal, though he does not correct for endogeneity or measurement errors. Fry and Lovell (2003) find that there is no need for bilingualism in the United States, but their estimates, also based on ordinary least squares, suffer from a downward bias. Galasi's (2003) IV estimates for Hungary are of the order of 0.10.

3. Instrumental variables quantile regression results

The results of the previous section give the returns to languages at the mean of the sample log wage distribution. It is also interesting to know how language returns vary at different points of the conditional distribution of log wages. This is precisely what quantile regression allows doing.[15] Moreover, since workers in the same quantile of the conditional wage distribution can be expected to have similar unobserved characteristics, quantile regression may help to control for unobserved heterogeneity. Increasing returns along quantiles could be a signal of a positive correlation between languages and unobserved heterogeneity.

However, although quantile regression could limit this endogeneity problem, it is not clear that it will completely remove it. Recent research by Chernozhukov and Hansen (2004, 2005, 2006) has extended quantile regression approach to deal explicitly with endogeneity. They propose a new instrumental variable quantile estimator that is naturally robust to weak identification and that is used in this research, with the same specification and instruments as in the regressions presented in Section 2. The results of the calculations for the same nine European countries are presented in graphical form for nineteen quantiles (0.05, 0.10, ..., 0.90, 0.95) in Figure 1. The shaded regions represent the 95 percent confidence intervals around the IV quantile regression estimates of the effect of the disenfranchisement rate on wages (plain curves). The horizontal discontinuous lines represent the GMM mean coefficients discussed in the previous section and the dotted straight lines delimit its 95 percent confidence interval. The effects of foreign language proficiency on earnings of both estimation methods are consistent, and lead to results that are very similar. Recall that the estimates can be interpreted as the percentage increase of wages for individuals who know a language other than the

[15] See Buchinsky (1998) and Koenker and Hallock (2001) for introductions to quantile regression.

domestic one and use it at their workplace (they still have to be converted using the disenfranchisement rate of the language in a specific country).

The following observations can be made. First, in most cases, and with the exception of Denmark and Greece, the estimated quantile effects are significantly different from zero at almost every quantile, showing that it pays to know (and use) foreign languages. They are, however, not significantly different from 0 for the lowest quantiles in Portugal and Spain. Second, we have found an important heterogeneity in the returns of foreign languages along the wage distribution for some countries suggesting that standard OLS or GMM estimations do not capture what is going on in all the quantiles.

In this sense, constant estimations along the wage distribution imply returns that are similar for all quantiles. This is so in Finland, France, Germany and Greece. For Italy the effect is stable for all quantiles, except the last ones, where the effect increases: for this country though language proficiency increases earnings in every point of the conditional distribution of wages, it is only at the very top of this distribution that the need for foreign languages gains in importance and is rewarded accordingly. We observe an increasing effect along the distribution in Austria, Denmark, Portugal and Spain. Hence, we find that increasing returns of foreign languages are not as common as expected under a strong positive correlation between unobserved ability and language returns. This can be interpreted as a new insight about the appropriateness of our instrument.

4. Conclusions

Our results show that in all nine countries, language proficiency has a positive effect on earnings.[16] The parameter of interest is affected by a substantial downward bias if estimated by ordinary least squares, which takes no account of time-varying measurement errors. This result suggests that the bias related to measurement errors can be more important than the ability bias. Dustmann and Van Soest (2001), Bleakey and Chin (2004) or Galasi (2003) observe comparable outcomes. The results reported in Table 5 show that mean returns can be very large and that there is heterogeneity between countries. Quantile estimation suggests that there is also significant heterogeneity within countries: the effects are larger in the upper deciles of the wage distribution for half of the countries analyzed which is not very surprising. This may, however, also be a consequence of the fact that we do not control enough for unobserved ability when using OLS or IV estimators.

The returns to languages that we isolate are much larger than those which are found in most studies on immigrants who acquire the language of the country to which they move. This is not surprising, since immigrants usually form small groups in countries where the largest part of citizens are fluent in the native language, and there is less market pressure to

[16] Recall that this is not so in Ireland, but there are not enough observations to make a precise statement.

pay higher wages to immigrants who learn and speak the national language, compared to those who do not learn it. Moreover, immigrants are often less skilled than national citizens, and as is shown in our quantile regressions, returns to languages are larger for the top quantiles of the earnings distribution. For lower paid jobs, it probably makes less of a difference whether the immigrant does or does not speak the domestic language.

The paper uses a market measure (the disenfranchisement rate of a language in every country) rather than a dummy (or a measure of language efficiency) for various languages. This makes it possible to "predict" returns for languages that are spoken by a small number of individuals, and for which returns can hardly be estimated directly, because the number of observations is too small. The link also outlines in a direct way that private returns decrease as disenfranchisement decreases, an interesting question since it isolates a tradeoff between public investment in language education and private returns. Public education will eventually crowd out private returns, as the number of speakers of a language should increase with more language education.

Given that English is the most widely known language, its returns are smaller than those that accrue to other, less known, languages. This may eventually lead to self-regulating dynamics in the learning behavior of citizens, in which English will yield to other languages whose returns are still larger once required at job but maybe a more complex search problem has to be faced by the worker to find a firm where this language is used (matching). This observation is consistent with Aydemir and Skuterud (2005, p. 642) who find that earnings of more recent immigration cohorts to Canada are deteriorating. They ascribe this to greater challenges in the Canadian labor market, but also to the more general economic trend that has reduced the earnings of native labor market entrants. This may alternatively be due to the fact that more immigrants know English than previously, which reduces the incentive for workers to learn English and for firms to reward the knowledge of English. This would become clearer if disenfranchisement rates were used.

Our findings can also be interpreted as showing that in Europe there is need to teach other languages than English.

References

Abbott, Michael and Charles Beach (1992), Immigrant earnings differentials in Canada: A more general specification of age and experience effects, *Empirical Economics* 17, 221-238.

Angrist, Joshua and Alan Krueger (1991), Does compulsory school attendance affect schooling and earnings?, *Quarterly Journal of Economics* 106, 979-1014.

Ashenfelter, Orley and Alan Krueger (1994), Estimates of the economic return to schooling from a new sample of twins, *American Economic Review* 84, 1157-1173.

Aydemir, Abdurrahman and Mikal Skuterud (2005), Explaining the deterioration entry earnings of Canada's immigrant cohorts, 1996-2000, *Canadian Journal of Economics* 38, 641-672.

Beenstock, Michael, Barry Chiswick and Gaston Repetto (2001), The effect of language distance and country of origin on immigrant language skills: Application to Israel, *International Migration* 39, 33-62.

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

Bleakey, Hoyt and Aimee Chin (2004), Language skills and earnings: Evidence from childhood immigrants, *The Review of Economics and Statistics* 86, 481-496.

Bratsberg, Bernt, James Ragan and Zafir Nasir (2002), The effect of naturalization on wage growth: A panel study of young male immigrants, *Journal of Labor Economics* 20, 568-597.

Buchinsky, Moshe (1998), Recent advances in quantile regression models: A practical guideline for empirical research, *The Journal of Human Resources* 33, 88-126.

Cattaneo, A. and R. Winkelmann (2003), Earning differentials between German and French speakers in Switzerland, Working Paper 0309, University of Zürich.

Chernozhukov, Victor and Christian Hansen (2004), Instrumental variable quantile regression, Discussion Paper, The Chicago University, Graduate School of Business.

Chernozhukov, Victor and Christian Hansen (2005), An IV model of quantile treatment effects, *Econometrica* 73, 245-262.

Chernozhukov, Victor and Christian Hansen (2006), Instrumental quantile regression inference for structural and treatment effect models, *Journal of Econometrics* 73, 245-261.

Chiswick, Barry (1998), Hebrew language usage: Determinants and effects on earnings among immigrants in Israel, *Journal of Population Economics* 15, 253-271.

Chiswick, Barry Yew Lee and Paul Miller (2003), Immigrants language skills: the Australian experinece in a longitudinal survey, *Annales d'Economie et de Statistique* 71-72, 97-129.

Chiswich, Barry and Paul Miller (1995), The endogeneity between language and earnings: international analyses, *Journal of Labor Economics* 11, 246-288.

Chiswich, Barry and Paul Miller (1996), Ethnic networks and language proficiency among immigrants, *Journal of Population Economics* 9, 19-35.

Chiswich, Barry and Paul Miller (2002), Immigrant learning: Language skills, linguistic concentrations and the business cycle, *Journal of Population Economics* 15, 31-57.

Crystal, David (2001), *A Dictionary of Language*, Chicago: Chicago University Press.

Dustmann, Christian (1994), Speaking fluency, writing fluency and earnings of migrants, *Journal of Population Economics* 7, 133-156.

Dustmann, Christian and Arthur Van Soest (2001), Language fluency and earnings: Estimators with miscalssified language indicators, *Review of Economics and Statistics* 83, 663-674.

Dustmann, Christian and Arthur Van Soest (2002), Language and the earnings of immigrants, *Industrial and Labor Relations Review* 55, 473-492.

Fry, Richard and B. Lindsay Lowell (2003), The value of bilingualism in the U.S. labor market, *Industrial and Labor Relations Review* 57, 128-140.

Galasi, Peter (2003), Estimating wage equations for Hungarian higher education graduates, manuscript.

Ginsburgh, V. and S. Weber (2005), Language disenfranchisement in the European Union, *Journal of Common Market Studies* 43, 273-286.

Hall, Alastair R. and Fernanda P.M. Peixe (2003), A consistent method for the selection of relevant instruments, *Econometric Reviews* 22, 269 – 287.

Hellerstein, Judith and David Neumark (2003), Ethnicity, language, and workplace segregation: Evidence from a new matched employer-employee data set, *Annales d'Economie et de Statistique* 71/72, 19-78.

INRA (2001), Eurobaromètre 54 Special, Les Européens et les langues, Février.

Klein, Carlo (2003), La valorisation des compétences linguistiques sur le marché du travail luxembourgeois, CEPS/INSTEAD Paper, Luxemburg.

Koenker, Roger and Kevin Hallock (2001), Quantile regression, *Journal of Economic Perspectives* 15, 143-156.

Leslie, D. and J. Lindley (2001), The impact of language ability on employment and earnings of Britain's ethnic communities, *Economica* 68, 587-606.

Shapiro, Daniel and Morton Stelcner (1997), Language earnings in Quebec: Trends over twenty years, 1970-1990, *Canadian Public Policy* 23, 115-140.

Williams, Donald (2005), The economic returns to multiple language usage in Europe, CEPS/INSTEAD Paper, Luxemburg.

Williams, Donald (2005), The economic returns to multiple language usage in Europe, CEPS/INSTEAD Paper, Luxemburg.

Appendix

The survey includes questions on the improvement of language capacities from one year to the next. An individual can chose between an improvement that gives him a good proficiency level (that is making him able to read complex information and to converse in most social contexts) or a poor to medium level (he became able to read basic information and to converse in routine situations). We can assume that individuals who have improved their language skills have constant unobserved ability and that the average unobserved ability is similar regardless of whether they have been able to use their new skills at job or not.

This makes it possible to select those individuals who declare improvements in their foreign language skills between 2000 and 2001, so that time-persistent unobserved ability is constant. Clearly, the number of observations is much smaller, and we had to pool observations over countries (the same countries as before, with the exception of Germany for which there is no information on language proficiency) but this sample will not suffer for ability bias. Due to the sample reduction, the same model as in the country by country estimation is used, except that country specific effects are included.

In Table A, we give the results for three equations. The first two equations are run on those individuals who declare having improved their language proficiency (to a good level, and to a poor or medium level). The third equation is run on the full sample, and is thus prone to a bias due to unobserved ability (but pools all countries, to make the results comparable to those of the two first equations). The results show that the parameters of the three equations are not significantly different. We ran a Wald-statistic to test whether the value of the disenfranchisement parameter γ is different from 0.45706 (the value for the whole sample in equation (3)) in equations (2) and (3). The values of the statistic are 1.03 and 0.35, respectively, well below any value of a χ^2 with 1 degree of freedom for any standard probability level. Hence, it seems that the ability bias is not significant when using the full sample since the results are not statistically different from those obtained with samples where unobserved ability is controlled.

Table A
Estimation Results
(Individuals who have improved their skills)

	Improvement of language skills		Total sample
	Good	Poor-medium	
	(1)	(2)	(3)
Disenfranchisement	0.32590	0.36184	0.45706
	(0.12936)	(0.16116)	(0.03169)
Higher Education	0.42001	0.40321	0.44834
	(0.03876)	(0.0371)	(0.01129)
Secondary Education	0.12164	0.11937	0.16988
	(0.03562)	(0.03172)	(0.00836)
Tenure	0.02275	0.01296	0.01044
	(0.00743)	(0.00601)	(0.00175)
Tenure sq.	-0.00035	0.00004	0.00008
	(0.00031)	(0.00025)	(0.00007)
Pot. experience	0.02161	0.02557	0.02538
	(0.00488)	(0.0037)	(0.00097)
Pot. Exper. sq.	-0.00032	-0.00040	-0.00043
	(0.00013)	(0.00009)	(0.00002)
Denmark	0.00341	-0.0161	0.01213
	(0.07979)	(0.09871)	(0.03791)
France	-0.04207	-0.09858	-0.05772
	(0.0791)	(0.09217)	(0.03491)
Italy	-0.32011	-0.39734	-0.37542
	(0.09153)	(0.10092)	(0.03535)
Greece	-0.80615	-0.75814	-0.73390
	(0.09308)	(0.1121)	(0.03609)
Spain	-0.41552	-0.50338	-0.43241
	(0.07418)	(0.09977)	(0.03467)
Portugal	-0.85117	-0.88052	-0.97398
	(0.08152)	(0.10191)	(0.03556)
Austria	-0.15039	-0.23348	-0.28041
	(0.08296)	(0.09908)	(0.03652)
Finland	-0.18617	-0.24772	-0.22207
	(0.09006)	(0.12627)	(0.03762)
Intercept	1.66459	1.70187	1.66383
	(0.07281)	(0.08079)	(0.0351)
R2	0.5849	0.5355	0.6321
No. of observations	807	1088	12933

Table 1
Overview of the data
(No. of observations)

	Austria	Belgium	Denmark	Finland	France	Germany	Greece	Italy	Portugal	Spain
No. of observations	1,166	864	1,029	928	1,649	2,383	1,304	2,214	2,242	2,401
Speakers of										
English	204	198	532	454	350	411	222	230	287	256
French	2	175	3	0	1,649	3	16	44	80	87
German	1,166	37	107	9	44	2,383	4	53	7	19
Italian	1	8	0	1	7	6	3	2,214	1	8
Spanish	0	2	2	0	14	4	0	4	28	0
Dutch	0	75	0	0	0	1	0	2	0	2401

Table 2
Disenfranchisement rates (per cent)

	English	French	German	Italian	Spanish
Austria	54	89	1	93	99
Denmark	25	95	63	100	98
Finland	39	99	93	100	99
France	58	0	92	95	85
Germany	46	84	3	99	98
Greece	53	88	88	92	95
Italy	61	71	96	1	97
Portugal	65	72	98	99	96
Spain	64	81	98	98	1

Source: Ginsburgh and Weber (2005, p. 279).

Table 3
Language switches between 2000 and 2001

(Number of individuals)

	Austria	Denmark	Finland	France	Germany	Greece	Italy	Portugal	Spain
No. of switches	61	250	167	128	37	149	159	195	210
No. of observations	1,166	1,029	928	1,649	2,383	1,304	2,214	2,242	2,401
Switches (in %)	6.1	24.3	18.0	7.8	1.6	11.4	7.2	8.7	8.8

These numbers should be read as follows. In Austria, for example, 61 out of 1,166 workers who responded to both surveys declared that the language they use at their place of work changed between 2000 and 2001.

Table 4
GMM Estimation results

	Austria	Denmark	Finland	France	Germany	Greece	Italy	Portugal	Spain
Second stage									
Disenfranchisement	0.283	0.189	0.503	0.505	0.495	0.275	0.292	0.471	0.608
	(0.093)	(0.103)	(0.095)	(0.058)	(0.056)	(0.099)	(0.053)	(0.076)	(0.118)
Higher Education	0.628	0.340	0.289	0.367	0.299	0.519	0.565	0.771	0.349
	(0.049)	(0.032)	(0.033)	(0.028)	(0.032)	(0.034)	(0.029)	(0.044)	(0.028)
Secondary Education	0.311	0.134	0.065	0.046	0.119	0.161	0.172	0.248	0.133
	(0.029)	(0.031)	(0.028)	(0.042)	(0.027)	(0.021)	(0.014)	(0.027)	(0.020)
Tenure	0.006	0.013	0.001	0.027	0.018	0.020	0.012	-0.005	0.014
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.003)	(0.004)	(0.004)
Tenure sq. (x100)	0.007	-0.047	0.028	-0.036	-0.012	-0.025	-0.013	0.067	0.019
	(0.023)	(0.023)	(0.028)	(0.020)	(0.018)	(0.020)	(0.013)	(0.018)	(0.018)
Pot. Experience	0.032	0.026	0.016	0.272	0.037	0.031	0.026	0.022	0.021
	(0.004)	(0.004)	(0.003	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
Pot. exper. sq. (x100)	-0.055	0.045	-0.025	-0.049	-0.072	-0.047	-0.045	-0.040	-0.035
	(0.008)	(0.008)	(0.007)	(0.007)	(0.008)	(0.006)	(0.005)	(0.003)	(0.004)
Intercept	1.256	1.832	1.676	1.533	1.385	0.817	1.305	0.740	1.247
	(0.039)	(0.051)	(0.040)	(0.030)	(0.052)	(0.033)	(0.023)	(0.024)	(0.217)
R-squared	0.388	0.227	0.274	0.362	0.241	0.487	0.427	0.402	0.403
First stage									
Disenfranch. lagged	0.628	0.488	0.611	0.635	0.914	0.487	0.616	0.552	0.449
	(0.042)	(0.036)	(0.033)	(0.029)	(0.012)	(0.040)	(0.034)	(0.032)	(0.043)
R-squared	0.449	0.246	0.479	0.457	0.887	0.303	0.337	0.412	0.222
No. of observations	1,166	1,029	928	1,649	2,383	1,304	2,214	2,242	2,401

Notes. The dependent variable is the (logged) wage rate. GMM standard deviations appear between brackets under the coefficients.

Table 5
Returns on languages

	Austria	Denmark	Finland	France	Germany	Greece	Italy	Portugal	Spain
English	0.15	0.05	0.20	0.29	0.23	0.15	0.18	0.31	0.39
French	*0.25*	*0.18*	*0.50*	0.00	0.42	0.24	0.21	0.34	0.49
German	0.00	0.17	*0.47*	0.46	0.00	*0.24*	0.28	*0.46*	0.60
Italian	*0.26*	*0.18*	*0.50*	*0.48*	*0.49*	*0.25*	0.00	*0.47*	*0.60*
Spanish	*0.28*	*0.18*	*0.50*	0.43	*0.49*	*0.26*	*0.28*	0.45	0.00
Dutch	*0.28*	*0.19*	*0.50*	*0.51*	*0.49*	*0.26*	*0.29*	*0.47*	*0.61*

Notes. Returns are obtained by multiplying the regression coefficient on disenfranchisement (first row) of Table 4 by the disenfranchisement rates appearing in Table 3. Since there are no immigrants in the samples, the return of the domestic is always set to 0, even if in Table 2, disenfranchisement rates can be different from 0 (for instance, there are 3 per cent of German citizens who claim they do not speak German). Returns for languages for which there are more than 10 speakers in the sample appear in bold.

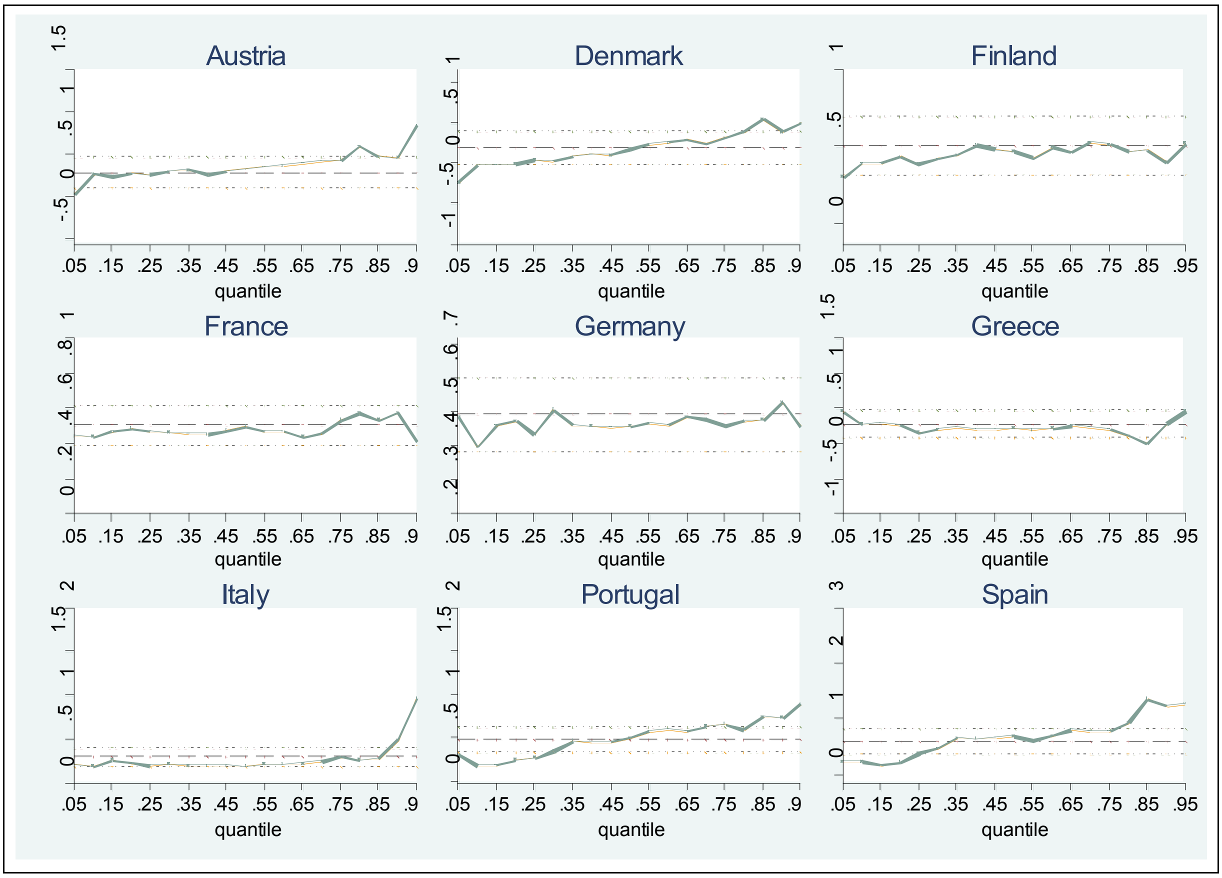
Table 6
Comparing OLS and GMM returns

	Austria	Denmark	Finland	France	Germany	Greece	Italy	Portugal	Spain
OLS									
Disenfranchisement	0.199	0.093	0.275	0.312	0.409	0.205	0.192	0.325	0.277
	(0.047)	(0.047)	(0.052)	(0.037)	(0.050)	(0.043)	(0.027)	(0.034)	(0.032)
R-squared	0.288	0.230	0.288	0.373	0.242	0.488	0.430	0.407	0.430
GMM									
Disenfranchisement	0.283	0.189	0.503	0.505	0.495	0.275	0.292	0.471	0.608
	(0.093)	(0.103)	(0.095)	(0.058)	(0.056)	(0.099)	(0.053)	(0.076)	(0.118)
R-squared	0.449	0.246	0.479	0.457	0.887	0.303	0.337	0.412	0.222

Figure 1

Quantile Regression Results

Disenfranchisement Coefficients and their Confidence Intervals



The shaded regions represent the 95 percent confidence intervals around the IV quantile regression estimates of the effect of the disenfranchisement rate on wages (plain curves). The horizontal discontinuous lines represent GMM mean coefficients and the dotted straight lines delimit their 95 percent confidence intervals.