The Economic Journal, **113** (*July*), 695–717. © Royal Economic Society 2003. Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

LANGUAGE PROFICIENCY AND LABOUR MARKET PERFORMANCE OF IMMIGRANTS IN THE UK*

Christian Dustmann and Francesca Fabbri

This paper uses two recent UK surveys to investigate the determinants of language proficiency and the effect of language on earnings and employment probabilities of non-white immigrants. We address the problem of endogenous choice of language acquisition and measurement error in language variables. Our results show that language acquisition, employment probabilities, as well as earnings differ widely across non-white immigrants, according to their ethnic origin. Language proficiency has a positive effect on employment probabilities, and lack of English fluency leads to earning losses.

According to the 2000 Labour Force Survey, immigrants (defined as individuals who are born outside UK) account for around 9% of the working age population of Britain. Immigrants are heavily concentrated in Metropolitan areas. In 2000, London contained around 9% of the total population of the UK, but more than 40% of all immigrants. The ethnic origin of immigrants in the UK is diverse, with the largest group being born elsewhere in the European Union, followed by immigrants from India, the Old Commonwealth, Pakistan, and Africa; see Dustmann *et al.* (2002) for more details.

A number of recent studies analyse various aspects of labour market behavoiur of ethnic minorities, and compare outcomes with those of the majority population; see Blackaby *et al.* (1994, 1997) and Clark and Drinkwater (2000). In much of this literature, however, no attempt is made to distinguish between immigrant and British born minorities. But many important questions are specifically related to first generation immigrants, who constitute a significant fraction of minorities in the UK. The 2000 Labour Force Survey shows that about 66% of ethnic minorities of working age were born abroad (Dustmann *et al.*, 2002).

There are few papers that investigate the economic assimilation of immigrants. The earliest study is Chiswick (1980), who uses data from the 1972 General Household Survey (GHS). His main finding is that, while white immigrants have very similar earnings patterns to native-born individuals, earnings of immigrants from ethnic minority groups are about 25% lower, other things the same. This gap is not decreasing with time of residence in the UK. In a more recent paper, Bell (1997) uses also data from the GHS, but he pools waves between 1973 to 1992. Like Chiswick, he finds that white immigrants are doing well. While white immigrants have an initial wage advantage, compared to native workers, immigrants from the West Indies and India have an earnings disadvantage, but wage differentials between this group and white natives decrease with the time spent in the UK. Shields

* We are grateful to Barry Chiswick, Hide Ichimura, Costas Meghir, Ian Preston, and Arthur van Soest for comments on earlier drafts of this paper, and to Barbara Sianesi for making her programme for propensity score matching available to us. We would like to thank David De Meza and three referees for comments that helped to improve the paper. The dataset used for this paper is not available to the public. If you need more information please contact the authors.

and Wheatley Price (1998) use data from the British Labour Force Survey. They emphasise the different assimilation patterns between foreign and native born minority individuals.

It may be in the interest of the host country to support the process of economic assimilation. To achieve this, it is important to understand the factors that determine the economic performance of minority immigrants. In this paper, we concentrate on one specific human capital factor, which is important not only for immigrants' economic assimilation, but also for their social integration: Proficiency in the host country language. Recent analyses for the US, Canada, Australia, Israel, and Germany show that fluency and literacy in the dominant host country language are important components for explaining immigrants' labour market success; see, Rivera-Batiz (1990), Chiswick (1991), Dustmann (1994), Chiswick and Miller (1995), Chiswick *et al.* (1997), and Berman *et al.* (2000). Work by Shields and Wheatley Price (2001) indicates that language is also positively related to occupational success of some immigrant groups in the UK.

We analyse the determinants of fluency and literacy in the host language for immigrants belonging to ethnic minority groups, and on how it relates to their labour market performance. We first investigate factors influencing the acquisition of the host country's language by the immigrant, such as education, age and years of residence in the host country. We distinguish between education received in the host and in the home countries.

We then analyse the extent to which language ability influences the labour market outcomes of immigrants. We focus on its effect on employment probabilities, and on the level of earnings. Estimates of language coefficients in straight forward regressions are bedeviled by two problems. First, as pointed out by Borjas (1994), the choice to acquire proficiency in a foreign language may be endogenous. Second, as stressed by Dustmann and van Soest (2001), language measures usually reported in survey data may suffer substantially from measurement error. The bias induced by these two problems points in opposite directions. We attempt to address both problems in this paper, and propose estimators which may help to reduce, or eliminate the bias. We combine an IV estimator that eliminates the bias due to measurement error with a matching estimator that addresses the problem of endogenous choice of language acquisition. Our results suggest that measurement error leads to a downward bias in the estimates of language on employment probabilities and earnings, and that the true effects are larger than OLS estimates reveal. These results are in line with findings for the other countries (Dustmann and van Soest, 2001).

Our best estimates suggest that fluency in English increases employment probabilities by about 22 percent points. This estimate is 5 percentage points higher than the OLS estimate. Furthermore, OLS estimates show that proficiency in English is associated with 18–20% higher earnings. Again, our estimator that takes account of both measurement error and endogenous selection indicates that effects are larger, but estimates are not significant, probably due to the small sample sizes.

We base our analysis on data from two UK surveys on ethnic minorities: the Fourth National Survey on Ethnic Minorities (FNSEM), which has been collected between 1993 and 1994, and the Family and Working Lives Survey (FWLS), which

has been collected between 1994 and 1995. Both data sets consist of two subsamples. The FWLS contains a main sample of the entire UK population and a 'boost' sample of individuals belonging to the ethnic minorities. The FNSEM contains a main sample of respondents belonging to ethnic minorities, and a reference sample of individuals belonging to the white majority population. Both surveys include questions on social and economic conditions of the interviewees, and measures on language proficiency. Information in the two data sets is complementary. For instance, while the FNSEM only reports spoken language proficiency, the FWLS contains information also about reading and writing skills. The FNSEM distinguishes between education acquired in home and host economy, information which is not available for the FWLS. Using two datasets allows us to conduct comparable analyses to check the robustness of the results obtained.

The data sources we use for this analysis are to our knowledge the only data sets for the UK that contain information about immigrants' language proficiency, as well as information on employment status and earnings. They are restricted to ethnic minority immigrants, and our results do therefore not necessarily generalise to the overall population of immigrants in the UK. Furthermore, results have to be evaluated subject to our ability to address the endogenous choice of language acquisition with the information available in the data, and the relatively small sample sizes, in particular for the earnings analysis. Despite these limitations, our analysis provides interesting insight into the relationship between language and economic outcomes for a large group of the UK's immigrant population.

The structure of the paper is as follows. Section 1 develops the estimation equations. Section 2 describes the data sets and gives some descriptive statistics. Section 3 investigates language determinants. Section 4 analyses how language proficiency affects employment probabilities, and earnings. Section 5 summarises the results and concludes.

1. Language and Labour Market Outcomes

The literature on migrants' earnings assimilation distinguishes between human capital which is specific to the host country, human capital which is specific to the home country, and human capital which is equally productive in both countries. Typically, immigrants enter the host country with skills which are only of limited use in the host economy, which results in an initial earnings disadvantage (Chiswick, 1978). After immigration, migrants transfer home country specific human capital into general or host country specific human capital, and acquire additional skills which are specific to the host country economy. The intensity of this process determines the speed of economic assimilation.

Language capital is an important component of host country human capital. It is also very specific to the host economy, since it is usually not transferable to the migrant's home economy. Standard human capital models may serve as a basis for formulating empirical specifications explaining the determinants of language capital (Dustmann, 1999). In such models, human capital is produced by investing time and other inputs. The cost of production equals forgone earnings, plus the cost of other input goods. A simple equilibrium condition states that investment in

human capital production is set such that the marginal cost equals the marginal benefit from the discounted future enhanced earnings potential. The production potential may differ across individuals according to their ability to acquire knowledge and it may depend on the stock of human capital acquired in the past. The benefit of any acquisition of host country specific human capital depends, in addition, on the length of the period over which it is productively put to use.

Investment in language capital should therefore depend on its potential future benefits, on the cost of acquisition, and on the individual's efficiency in producing it. Chiswick and Miller (1995) provide an extensive discussion on the variables which represent these factors. Variables which measure the immigrant's efficiency in acquiring language capital are the level of education upon immigration and the age at immigration (since the learning potential may deteriorate over the life cycle). The cost of acquiring the host country language depends on the distance of the migrant's mother tongue from the dominant majority language, which may be captured by country of origin dummies. Clearly, this last variable picks up a variety of other factors which affect language proficiency, like different degrees of immigrant selection across countries (Borjas, 1985, 1987). Assuming that all migrations are permanent, the time period over which language capital is productive depends on the migrant's age at entry. Accordingly, those who migrate at younger age should have a higher incentive to acquire language capital. Its acquisition may, in addition, depend on the extent to which individuals are exposed to the language of the majority population. As noted by Chiswick and Miller (1995), a variable which measures exposure is the time of residence abroad.

1.1. *Language, Earnings, and Employment Probabilities*

When analysing the effect of language on labour market outcomes, two problems may occur. First, the choice of learning the host country language may be endogenous, and related to variables which affect outcomes. This may lead to an upward bias of estimated language effects on economic outcomes. Second, unsystematic measurement error may lead to a downward bias of the effect of language on earnings. Numbers presented in Dustmann and van Soest (2001) on repeated language information for the same individual suggest that measurement error is substantial in self-reported language measures. In fact, in their data, more than half of the within individual variation in language responses is due to measurement error. Their results suggest that the downward bias induced by measurement error overcompensates the upward bias induced by unobserved heterogeneity.

To give a causal interpretation to the language coefficient, we need to deal with both sources of bias. We first discuss the problem of the endogenous choice of language acquisition. Assume for the moment that the language variable is measured without error. Then the problem is that those individuals who have chosen to obtain proficiency in the English language may differ from those individuals who have chosen not to do so. If these differences affect outcomes (in our case, employment or earnings) other than through language, a comparison in outcomes of the two groups does not produce an unbiased estimate of the causal effect of language proficiency.

We define the parameter we would like to obtain as the difference in outcomes for an individual of being proficient and non-proficient, after having made the choice of acquiring language proficiency.[1] Denoting these two potential outcomes by y_i^1 and y_i^0, and proficiency in English by $l_i = 1$, where i is an index for individuals, this parameter is given by

$$\mathrm{E}(y_i^1 - y_i^0 | l_i = 1).$$

This mean effect of language proficiency on outcomes for those who have decided to learn the foreign language is often referred to as the effect of 'treatment on the treated'; see Heckman *et al.* (1998). The problem we face in retrieving this parameter is that we do not observe individuals who decided to learn the host country language, but then refrained from doing so. In other words, the counterfactual $\mathrm{E}(y_i^0 | l_i = 1)$ is not observed. What we observe instead is $\mathrm{E}(y_i^0 | l_i = 0)$. If individuals who have, and who have not chosen to learn the language differ in characteristics related to outcomes, $\mathrm{E}(y_i^1 - y_i^0 | l_i = 1) \neq \mathrm{E}(y_i^1 | l_i = 1) - \mathrm{E}(y_i^0 | l_i = 0)$.

To estimate the mean effect of language on outcomes for those who have chosen to learn the language, we use a matching type approach. Suppose that we observe a vector of conditioning variables $\mathbf{x}_i$, sufficient to control for the endogenous choice of learning the English language. Then the expectation of the outcome with no language proficiency is conditionally independent of the decision to learn the language, i.e. $\mathrm{E}(y_i^0 | \mathbf{x}_i, l_i = 0) = \mathrm{E}(y_i^0 | \mathbf{x}_i, l_i = 1)$. Under this conditional independence assumption, we can use the outcome of those who are not proficient in the English language to estimate the counterfactual outcome of those who are proficient, were they not proficient. The parameter of interest is then given by $\mathrm{E}(y_i^1 | l_i = 1, \mathbf{x}_i) - \mathrm{E}(y_i^0 | l_i = 0, \mathbf{x}_i)$, which can be obtained from the data.

If $\mathbf{x}_i$ is multi-dimensional, this amounts to comparing individuals with the same cell distribution in terms of the variables in $\mathbf{x}_i$. This requires large data sets, and discretisation of continuous variables in $\mathbf{x}$. Rosenbaum and Rubin (1983) show that, if the conditional independence assumption is fulfilled, then it suffices to match on the propensity score $\mathrm{P}(l_i = 1 | \mathbf{x}_i) = P(\mathbf{x}_i)$ (the probability of being proficient in English, conditional on characteristics $\mathbf{x}_i$), which reduces the matching index to one dimension.

It is important to ensure that individuals are only matched for those $\mathbf{x}_i$ commonly observed for proficient, and non-proficient individuals (i.e. who have a common support in $\mathbf{x}$). If, for instance, there are values of $\mathbf{x}_i$ where only proficient individuals are observed – in other words, $P(\mathbf{x}_i) = 1$ for some values of $\mathbf{x}_i$ – the conditional expectation of $\mathrm{E}(y_i^0 | l_i = 0, \mathbf{x}_i)$ is not defined. Heckman *et al.* (1997) show that, if the common support condition is not fulfilled, then the matching approach may lead to seriously biased estimates.

We use a propensity score estimator, which ensures that the support conditions are fulfilled. We estimate the propensity score for being proficient in the English

[1] An alternative parameter of interest is the difference in outcomes of being proficient and non-proficient in the English language for individuals who have chosen not to learn the language. See Dearden *et al.* (2000) for a discussion of the two parameters.

language using a simple logit model. We estimate the conditional expectation of the counterfactual using a Gaussian kernel, and match observations by nearest neighbour matching, based on the propensity score. We disregard individuals for which the absolute difference in the propensity score to the nearest neighbour in the control sample is not small enough. We then compute the mean difference between the treatment group and the constructed counterfactual. We estimate $\gamma^M = \int E\{y_i^1 - \mathrm{E}[y_i^0 \,|P(\mathbf{x}_i),\, l_i = 0] \,|\, l_i = 1\}\, dF[P(\mathbf{x})]$, where $\mathrm{E}[y_i^0 \,|P(\mathbf{x}_i),\, l_i = 0]$ is estimated using a Gaussian Kernel on those who are not proficient in the English language. Finally, we compute standard errors by bootstrap, using 500 repetitions.

A second problem we face is that there is measurement error in the self-reported language indicator. To address the measurement error problem, we use a two stage approach, which is based on the following idea. Suppose we had an instrument I_i, which has the properties that (i) it is independent of the outcome, conditional on x_i and l_i and (ii) it explains variation in l_i (in other words, $\mathrm{E}(l_i|I_i = r)$ is a non-trivial function of r, where r is in the support of I). These conditions correspond to the rank and order conditions for instrumental variable estimation. Let the instrument I_i be binary (in our case, another measure of language). Then an estimator which corrects for individual heterogeneity (using the matching approach) and measurement error (using an IV argument) is given by

$$\gamma^{MI} = \frac{\mathrm{E}(y_i|I_i = 1, \mathbf{x}_i) - \mathrm{E}(y_i|I_i = 0, \mathbf{x}_i)}{\mathrm{Prob}(\tilde{l}_i = 1|I_i = 1, \mathrm{x}_i) - \mathrm{Prob}(\tilde{l}_i = 1|I_i = 0, \mathrm{x}_i)}, \tag{1}$$

where $\tilde{l}_i$ is the measured binary language variable. To estimate this parameter, we proceed in two stages. In the first stage, we compute the numerator of (1) by propensity score matching, using the binary instrument I_i (which is the interview language) instead of the language variable. In a second step, we re-scale this parameter. We compute the denominator as the difference in the predicted probabilities of our language measure (using a linear probability model) for the two outcomes of the instrument.[2] We then compute the ratio of the two to obtain an estimate of the effect of language on outcomes, which takes account of both endogenous choice and measurement error. To compute the standard errors, we use bootstrapping.

The matching approach is based on the idea that the observable characteristics are sufficient to explain any relationship the choice of learning the language has on the outcome if non-proficient in English. In both data sets, we observe individual specific characteristics (like education, age, origin) and minority concentration in the area. Education should be correlated with otherwise unobserved determinants of the choice to acquiring language proficiency, like innate ability. In the two data sets, some information about family and household characteristics

[2] The intuition is as follows. The numerator is the change in the outcome variable if the instrument switches from zero to one; the denominator is the change in the probability of being proficient if the instrument switches from zero to one. It is easy to show that the expression in the denominator is equal to the change in the probability of being proficient in the true language measure if the instrument switches from zero to one, as long as the instrument is not correlated with the measurement error. The ratio of the two is then the change in the outcome variable if the true language variable switches from zero to one. See Heckman (1997) for a discussion of similar estimators.

is available. For the FNSEM, we include marital status, number of children, and partner characteristics. In the FWLS, we only observe marital status and number of children, but we have information on some self-assessed abilities, like mental arithmetic, and finding an address on a map.

2. The Data

The Family and Working Lives Survey (FWLS) was collected in 1994 and 1995. It is a retrospective survey on adults aged between 16 and 69, including 9,000 respondents and their partners. It contains a 'boost' sample of about 2,000 individuals belonging to four racial minority groups: Black Caribbeans, Indians, Pakistanis and Bangladeshis. The data provide information on earnings, education, nationality, language skills and other background characteristics. Of the 2,388 people forming the minority sample in the main and 'boost' sample, 68% (1,639) are foreign born.

The Fourth National Survey on Ethnic Minorities (FNSEM) is also a cross- sectional survey, which has been carried out between 1993 and 1994. Individuals included are aged 16 or more, and of Caribbean, Indian, Pakistani, Bangladeshi, and/or Chinese origin. There are 5,196 observations in the minority sample, and 2,867 observations in the independent comparison sample of white individuals. Similarly to the FWLS, more than 77% (4,019) of the individuals in the ethnic minority sample are foreign born.

The FWLS identifies the ward where the individual lives.[3] It is therefore possible to match this data set with the 1991 Population Census to construct a variable on the ethnic concentration on ward level.[4] The FNSEM does not contain geographical identifiers; therefore, matching with the Census data is not possible. However, it contains grouped information on ethnic concentration at ward level, obtained by the authors of the survey from the 1991 Census.

Both data sets provide information on earnings. The FWLS reports weekly gross (before tax) earnings, while the FNSEM reports grouped gross weekly earnings. Both data sets report the main activity of the individual (e.g. full-time or part-time paid work, full-time education, unemployed, etc.).

The sample design of the two surveys differs substantially. The ethnic minority sample of the FWLS was selected by screening addresses in areas where the ethnic minority population, according to the 1991 Census, was more than 3% of the local population. The selection in the FNSEM was more complex, considering wards with any percentage of ethnic minorities on the population and oversampling Bangladeshis to obtain a sufficient sample size. For more details, see Appendix 1 in Modood and Berthoud (1997), and Smith and Prior (1996).

Table 1 shows the percentage of immigrants belonging to ethnic minorities with respect to the overall population in the UK (column 1), and the ethnic composition within the group of ethnic immigrants. Numbers are based on the 1991

[3] In the UK, a ward is the smallest geographical area identified in the Population Census, the mean population within a ward is 5,459 individuals, and the median is 4,518.

[4] We define ethnic concentration as the ratio of the number of individuals belonging to ethnic minorities over the total population living in the ward. See footnote to Table 3 for details.

Table 1

Ethnic Immigrant Composition in the UK (Census 1991)

	Immigrants % of UK Pop.	Ethnic composition	Ethnic composition without Africans
Caribbean	0.56	18.19	23.41
Indian	0.84	27.57	35.49
African	0.68	22.31	–
Bangladeshi	0.22	7.09	9.13
Pakistani	0.47	15.46	18.89
South East Asian	0.29	9.37	12.06
Total	3.06	100	100

Table 2

Ethnic Immigrant Composition in Survey Data

	FWLS		FNSEM		
	No.	%	No.	%	%
Black Caribbean	265	16.17	698	18.20	17.37
Indian	314	19.16	971	25.32	24.17
Afro-Asian	123	7.50	656	17.11	16.32
Bangladeshi	512	31.24	550	14.34	13.68
Pakistani	425	25.93	960	25.05	23.89
Chinese	–	–	184	–	4.58
Total	1,639	100	4,019	100	100

Census. Table 2 gives the ethnic composition of the two surveys. Both surveys do not include Black African immigrants, and the FWLS does not include the Chinese minority. In the last column of Table 1, we report respective numbers in the census, excluding Africans. Comparing the two Tables, it appears that both surveys tend to oversample the South Asian groups (Indians, Pakistanis and Bangladeshis). Also, the two surveys differ in the ethnic composition of the respondents: Bangladeshis amount to 31% in the FWLS and 14% in the FNSEM, Indians to 19% in FWLS and 24% in the FNSEM and African Asians to 8% in the FWLS and 17% in the FNSEM.

Both surveys contain information on language. In the FWLS, language ability is self-assessed. The individual is first asked whether English is his/her mother tongue. If not, the individual is asked to self-assess proficiency in speaking, reading and writing English on a 5 point scale. The FNSEM contains two variables which are related to language proficiency: first, the interviewer's evaluation of the individual's spoken language ability, on a 4 point scale. Second, information about what fraction of the interview was held in English. In all areas with a minority density above 0.5% (which includes 97% of the sample individuals), there was an initial screening interview with the interviewee. In the case of poor fluency, the interviewers were chosen to be fluent in the language of the respondents. During the interview, interviewers decided about the extent to which English could be used in the interview, and we have information as to whether the interview was

held wholly in English, partly in English, or wholly in the individual's mother tongue.

In Table A1 we display the responses to self-assessed (FWLS) or interviewer assessed (FNSEM) language questions for the two data sets, broken down according to ethnic origin. The general pattern is similar for the two data sets.

For the empirical analysis, we re-defined the language indicators in the two surveys as dichotomous variables. For the FWLS, the variable assumes the value 1 if the individual reports language fluency or literacy as 'quite well' or 'very well', or reports English as a first language. For the FNSEM, it is equal to 1 if individuals fall in the categories 'fairly' or 'fluently'. We use the information on the interview language in the FNSEM as an instrument for measurement error. Our instrument is equal to one if the interview was done in English only.

Table 3 explains the variable used for the analysis, and presents summary statistics. The mean values on language indicate that the percentage of individuals who speak the English language at least fairly (or quite well) is very similar in the two samples. Percentages for reading and writing in English (available in the FWLS) are slightly lower.

About 51% (FWLS) and 56% (FNSEM) of the sample populations are in the labour force. Of these, 70% (FWLS) and 75% (FNSEM) are employed. These numbers are remarkably similar for the two data sets.

Table 3

Variables Description and Sample Characteristics

	FWLS		FNSEM		
Variable	Mean	S.D	Mean	S.D	Description
speak	0.709	0.454	0.691	0.462	Dummy = 1 if spoken English is good or very good
read	0.671	0.469	–	–	Dummy = 1 if read English is good or very good
write	0.641	0.479	–	–	Dummy = 1 if written English is good or very good
LabFo	0.511	0.500	0.559	0.469	Dummy = 1 if in labour force
empl	0.703	0.457	0.749	0.433	Dummy = 1 if employed (conditional on *LabFo* = 1)
Wgearn	239.175	432.809	240.049	–	Weekly gross earnings
Sex	0.468	0.499	0.505	0.500	Dummy = 1 if male
age	38.347	13.588	42.604	14.407	Age
yearstay	20.404	10.313	21.367	10.001	Years of residence in the UK
married	0.726	0.446	0.776	0.417	Dummy = 1 if married
nchild	1.937	1.793	1.654	1.761	Number of children in household
*Degree**	0.072	0.258	0.127	0.333	Dummy = 1 if university degree
*Alev**	0.129	0.335	0.109	0.312	Dummy = 1 if A Levels or higher vocational qualification
*OlevCSE**	0.231	0.422	0.230	0.421	Dummy = 1 if O Levels, medium or lower vocational qualification
noqual	0.568	0.495	0.533	0.499	Dummy = 1 if no qualification
ethcon	0.168	0.153	0.166	0.189	Ward ethnic minority concentration**
carib	0.1620	0.369	0.178	0.383	Dummy = 1 if Black Caribbean
indian	0.186	0.389	0.245	0.429	Dummy = 1 if Indian
afroas	0.0838	0.277	0.169	0.375	Dummy = 1 if African Asian
pakista	0.255	0.436	0.218	0.413	Dummy = 1 if Pakistani
chinese	–	–	0.048	0.214	Dummy = 1 if Chinese
bangla	0.318	0.466	0.142	0.349	Dummy = 1 if Bangladeshi

*Definitions follow Dearden (1999). **Defined as the ratio of ethnic minority individuals over the total population.

The mean value of weekly wages in the FLWS is £239.17, considering both part and full-time workers. Mean weekly wages are reported in the FNSEM as a grouped variable. The mean weekly gross wage is £240, which is similar to the mean wage in the FWLS.[5]

The average education level is slightly higher in the FNSEM than in the FWLS, with 12.7% graduates in the former sample, and only 7.2% in the later sample. Furthermore, there is a slightly higher percentage of individuals with no qualification in the FWLS (56.8%) than in the FNSEM (53.3%).[6]

The average ethnic minority concentration at ward level amounts, in both samples, to more than 16% (the average ward concentration in the FNSEM is obtained by taking the average of the mid-point values of the grouped variable, since the information is available only in intervals). The considerable difference in the sample designs is reflected only by the larger standard deviation indicated in the FNSEM.

In Table A2, we break down means of the age at immigration, year of immigration and the age of the various ethnic groups. In the FWLS, individuals are on average four years younger than in the FNSEM and have immigrated at a younger age. The immigration patterns for the various ethnic groups are similar in both data sets and correspond to the migration patterns indicated by Bell (1997) and Hatton and Wheatley Price (1999): Black Caribbeans arrivals are concentrated in the late 1950s and early 1960s, whereas Indians, African Asians and Pakistanis arrived mainly during the 1970s and Bangladeshis towards the end of the 1970s. Consistent with their shorter stay, Bangladeshis are the youngest group, whereas Black Caribbeans are the oldest on average.

3. Language Determinants

After eliminating all the observations with missing values in the variables of interest, we are left with 1,589 observations in the FWLS sample, and 3,732 observations in the FNSEM sample.

Table 4 reports coefficient estimates and robust standard errors from linear probability models, where the indicator variable equals one if the individual is proficient in the respective language component.[7] Comparing results on spoken language for the two data sets shows that the signs of regressors are equal for both samples in most cases, and the sizes of the coefficients are likewise similar (although the coding of the fluency variables differs slightly). Males have a significantly higher probability to be fluent in the majority language. The effect of age (which corresponds to the effect of age at entry, since we condition on years of residence) is negative and strongly significant. Years of residence has the expected

[5] Information on earnings is grouped in the FNSEM. To obtain this number, we estimate a grouped regression model on a constant, and compute the mean of the prediction (Stewart, 1983).

[6] We construct the education variables, following a classification by Dearden (1999): the variable *Degree* defines University degree or post-graduate diploma; the variable *Alev* stands for A-Levels or higher vocational degree; the variable *OlevCSE* includes O-levels, middle or lower vocational degrees and miscellaneous qualifications.

[7] We have also estimated probit models. Marginal effects, evaluated at the sample means, are almost identical to the coefficients we report in the Tables.

Table 4

Language Determinants, Linear Probability Models

	FWLS						FNSEM			
	Speaking		Reading		Writing		Speaking			
							All Qualifications		UK/nonUK Q	
Variable	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE
Const	0.616**	0.083	0.639**	0.084	0.640**	0.085	0.778**	0.053	0.872**	0.055
male	0.105**	0.019	0.109**	0.019	0.082**	0.019	0.144**	0.012	0.152**	0.012
age	−0.013**	0.004	−0.014**	0.004	−0.018**	0.004	−0.024**	0.002	−0.030**	0.002
$age^2/100$	0.010*	0.005	0.010*	0.005	0.016**	0.005	0.014**	0.002	0.019**	0.002
yearstay	0.021**	0.003	0.012**	0.004	0.012**	0.004	0.023**	0.002	0.027**	0.002
$yearst^2/100$	−0.036**	0.010	−0.014	0.010	−0.018**	0.010	−0.027**	0.005	−0.034**	0.006
Degree	0.308**	0.037	0.415**	0.038	0.457**	0.038	0.400**	0.019	–	–
Alev	0.303**	0.028	0.362**	0.029	0.421**	0.029	0.275**	0.019	–	–
OlevCSE	0.299**	0.023	0.337**	0.023	0.380**	0.023	0.223**	0.015	–	–
Edegree	–	–	–	–	–	–	–	–	–	–
EAlev	–	–	–	–	–	–	–	–	0.190**	0.023
EOlevCSE	–	–	–	–	–	–	–	–	0.182**	0.019
Fdegree	–	–	–	–	–	–	–	–	0.461**	0.023
FAlev	–	–	–	–	–	–	–	–	0.234**	0.029
FOlevCSE	–	–	–	–	–	–	–	–	0.195**	0.018
married	−0.047*	0.023	−0.053*	0.024	−0.039	0.024	0.004	0.015	0.006	0.016
nchild	−0.016**	0.006	−0.012*	0.006	−0.018**	0.006	−0.005	0.003	−0.006*	0.003
indian	0.249**	0.030	0.230**	0.030	0.223**	0.030	0.089**	0.021	0.087**	0.021
afroas	0.241**	0.037	0.236**	0.038	0.215**	0.038	0.232**	0.022	0.258**	0.023
pakista	0.137**	0.025	0.075**	0.025	0.074**	0.025	−0.021	0.019	−0.019	0.020
carib	0.373**	0.036	0.396**	0.037	0.435**	0.037	0.454**	0.024	0.482**	0.025
chinese	–	–	–	–	–	–	0.071*	0.031	0.069*	0.034
ethcon	−0.468**	0.091	−0.316**	0.093	−0.181	0.093	−0.208**	0.031	−0.215**	0.032
No. of Obs.	1,589		1,589		1,589		3,732		3,552	
Obs. Prob.	0.710		0.646		0.641		0.691		0.675	

Base Category: no educational qualification, Bangladeshi. Ethnic concentration for FNSEM at midpoints. Robust standard errors are reported. *Significant at 5% level. **Significant at 1% level.

positive effect, which decreases with time in the host country. All these results are consistent with findings for other countries. For the FWLS, the effect of these variables is similar for all three components of language capital.

The effect of the education variables is quite strong for fluency (the comparison group are individuals who report to have no qualification). For instance, for the FWLS (FNSEM) individuals with O-levels or equivalent have a 29 (22) percentage points higher probability of being fluent in English.

Speaking fluency may largely be acquired by exposure to the host country language, while writing and reading in a foreign language is a skill which is more difficult to obtain. Acquisition requires a more systematic way of learning, and the general level of schooling obtained may enhance the efficiency of acquiring this component of language capital. This is reflected by our results, which indicate that educational background variables have larger coefficients for reading and writing skills.

Education may be partly obtained in the host country. Since those who wish to enter the educational system in the UK are likely to have acquired some language skills, this leads to a classical simultaneity bias.

The FNSEM allows us to distinguish between education obtained in the UK and abroad. We have re-estimated the language equation, distinguishing between education obtained overseas, and in the UK. Results are reported in the last column of Table 4. We denote by *F* educational achievements obtained abroad, and by *E* educational achievements obtained in the UK.[8] The effect of overseas qualifications on language fluency is very similar to the effect of education obtained in the UK.

The variable *nchild* measures the number of children in the household. Chiswick and Miller (1995) suggest that children may have counteracting effects on language: first, they may act as a translator between the parent and the English speaking community (thus reducing incentives to learn the foreign language). Second, they may enhance exposure to the majority population by forcing the parent to cope with institutional matters, like school and parents of native friends of children. Our results indicate that children coefficients are negative for both data sets and for all language components.[9]

There are large differences in the level of language proficiency across different ethnic groups. Results of both data sets indicate that Bangladeshis, the base group, are dominated by nearly all other ethnic groups, except for Pakistanis in the FNSEM.

The variable *ethcon* measures ethnic concentration at ward level. It is strongly associated with language proficiency for both data sets. Results from the FWLS indicate that an increase in the ethnic density by 1 percentage point is associated with a 0.47 percentage point decrease in the probability to be fluent in the dominant language. The negative association with reading and writing skills is slightly smaller. Results from the FNSEM also indicate a negative association, but the size of the coefficient is only half as large as that for the FWLS. These results are in line with findings for the US, Canada and Israel (Chiswick, 1994; Chiswick and Miller, 1995).

4. Language and Economic Outcomes

4.1. *Employment Probabilities*

Language proficiency is likely to be a decisive factor in determining employment probabilities. Language may help to acquire information about optimal job search strategies. Migrants who are not sufficiently proficient in the dominant language may have difficulties in convincing prospective employers of their qualifications. Also, many jobs, for instance in the service sector, require communication skills.

[8] The variable *Edegree* predicts outcomes perfectly. Estimations are performed on the sample of non-degree holders.

[9] We have also estimated models where we interact number of children with gender. The children variable is positive (though insignificant) for males, but negative (and significant for the FWLS data) for females.

Likewise, literacy in the dominant language is a crucial prerequisite for many unskilled occupations.

To understand the association between employment probabilities and language, we consider individuals who are in the labour force, and we distinguish between those who are in work, and those who are not employed, but who are actively seeking a job.[10] Our samples consist of 839 individuals for the FWLS, and 2,100 individuals for the FNSEM. Our dependent variable, *EMPL*, takes the value 0 if the individual is unemployed and seeking a job or claiming benefits, and the value 1 if the individual works full or part-time. Explanatory variables are the demographic and human capital characteristics available in the two data sets, including a dummy variable for the level of language proficiency. The results are reported in Table 5. For the FWLS, we report results conditioning on fluency only, and on fluency and written literacy.

Most coefficient estimates for the two data sets are very similar. Males have a significantly lower probability of being employed (13 percentage points in the FWLS and 8 percentage points in the FNSEM). Being married increases employment probabilities by about 18 (17) percentage points. On the other hand, having children influences the employment probability negatively. These effects are consistent with evidence for British (male) natives (Nickell, 1980).

For the FWLS, education coefficients are mostly insignificant. For the FNSEM, education coefficients are significant and of the expected order of magnitude. In the last columns of Table 5, we show regressions which distinguish between education levels acquired in the UK and in the home country. The coefficients on the UK educational degrees seem slightly larger than the coefficients on education acquired at home. However, we can not reject the null hypothesis that the coefficients are equal (neither in isolation, nor jointly).

Age is positively associated with employment probabilities, and the age profile is concave. Conditional on age, the time of residence in the UK does not have a significant effect on employment probabilities, for both the FWLS and the FNSEM. Indians, Afro-Asians and Chinese have higher probabilities of being employed than Pakistanis and Bangladeshis. Again, Bangladeshis seem to be the most disadvantaged group.

The coefficients on the language variables are quite large, and similar for the two data sets. English fluency is associated with a 15 (17) percentage point higher employment probability, using the FWLS (FNSEM) data. The coefficients are highly significant.

The FWLS data distinguish between speaking, writing and reading abilities – information which is not available in most datasets on migrants' language abilities. One may argue that proficiency in the spoken language alone is not sufficient to affect labour market outcomes but that writing skills are likewise needed. The positive coefficient of the fluency variable may then simply reflect the correlation between these two components of language capital. To investigate this point, we

[10] This follows the ILO definition of unemployment. According to the ILO definition, people are considered as unemployed if aged 15 years or older, without work, but available to start within the next two weeks and have actively sought employment at some time during the previous four weeks.

Table 5

Employment Probabilities, Linear Probability Models

	FWLS						FNSEM			
	1		2		3		4		5	
							All Qualifications		UK/nonUK Q	
Variable	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE
Const	−0.052	0.169	−0.082	0.169	−0.087	0.169	0.101	0.116	0.105	0.118
male	−0.128**	0.034	−0.123**	0.034	−0.125**	0.034	−0.080**	0.019	−0.079**	0.019
married	0.175**	0.042	0.176**	0.041	0.178**	0.042	0.167**	0.025	0.168**	0.025
nchild	−0.035**	0.011	−0.034**	0.011	−0.034**	0.011	−0.026**	0.006	−0.026**	0.006
degree	0.047	0.053	0.019	0.055	0.018	0.055	0.107**	0.026	–	–
Alev	0.008	0.045	−0.016	0.047	−0.017	0.047	0.121**	0.027	–	–
OlevCSE	−0.064	0.039	−0.084*	0.040	−0.086*	0.040	0.071**	0.022	–	–
Edegree	–	–	–	–	–	–	–	–	0.103**	0.034
EAlev	–	–	–	–	–	–	–	–	0.116**	0.030
EOlevCSE	–	–	–	–	–	–	–	–	0.069**	0.025
Fdegree	–	–	–	–	–	–	–	–	0.082**	0.032
FAlev	–	–	–	–	–	–	–	–	0.067	0.040
FOlevCSE	–	–	–	–	–	–	–	–	0.052*	0.025
age	0.029**	0.009	0.030**	0.009	0.030**	0.009	0.016**	0.006	0.016**	0.006
$age^2/100$	−0.039**	0.012	−0.040**	0.011	−0.040**	0.011	−0.024**	0.007	−0.024**	0.007
yearstay	0.002	0.007	0.004	0.006	0.003	0.007	0.003	0.004	0.003	0.004
$yearst^2/100$	0.001	0.017	−0.000	0.017	−0.001	0.017	−0.004	0.010	−0.003	0.010
black	0.105	0.059	0.094	0.059	0.089	0.059	0.126**	0.039	0.127**	0.039
afroas	0.128*	0.057	0.131*	0.057	0.125*	0.057	0.182**	0.035	0.183**	0.035
indian	0.172**	0.049	0.173**	0.048	0.166**	0.049	0.177**	0.033	0.183**	0.033
pakista	0.064	0.045	0.071	0.045	0.066	0.045	0.024	0.033	0.029	0.033
chinese	–	–	–	–	–	–	0.250**	0.046	0.243**	0.047
speak	0.147**	0.046	–	–	0.049	0.062	0.171**	0.025	0.169**	0.025
write	–	–	0.164**	0.042	0.133*	0.057	–	–	–	–
No. of Obs.	839		839		839		2,100		2,100	

Base Category: no educational qualification, Bangladeshi. Robust standard errors are reported. *Significant at 5% level. **Significant at 1% level.

have included an indicator for writing abilities (columns 2), and both speaking and writing variables (columns 3). The effect of writing proficiency (unconditional on fluency) is slightly higher. When including both indicator variables, we find that writing abilities are associated with a 13 percentage point increase in employment probabilities, while speaking ability alone increases this probability by only 5 percentage points. The latter effect is not significant. This suggests that literacy in the dominant majority language, in addition to fluency, is important to obtain a job.

4.2. *Employment, Endogenous Choice and Measurement Error*

The above results suggest that language proficiency has a positive impact on employment probabilities. As we discussed above, however, the estimated coefficients may be seriously biased due to endogenous choice and measurement error. Furthermore, the effect of language on employment may be different for males and females. In this section, we address these issues. We estimate different models, addressing both these problems, and using the pooled sample, and males and females separately. We report the results in Table 6.

In the first row, we replicate our OLS results (based on the same specification as in Table 5), where we also report estimates for males and females separately. For the FNSEM data, the language coefficient is very similar for males and females, and significantly different from zero for both groups. For the FWLS, the coefficient for males is slightly larger than the coefficient for the pooled sample, while the coefficient for females is practically zero.

The second row reports results using the propensity score matching estimator, as we have explained in Section 1. Coefficients decrease slightly, which is compatible with unobserved ability being still present in the simple regression in row 1.

In the last row, we report results from implementing the two stage estimator which takes account of measurement error (see (1) above). Coefficient estimates increase quite substantially. The results suggest that measurement error in the language variable leads to a substantial downward bias in estimated parameters.

Table 6

Employment and Language

		FNSEM			FWLS		
Specification		All	Males	Females	All	Males	Females
1: OLS	Coeff	0.170	0.166	0.172	0.147	0.190	−0.007
	StdE	0.025	0.024	0.041	0.046	0.037	0.070
2: Prop. Match.	Coeff	0.102	0.102	0.133	0.100	0.112	−0.140
	StdE	0.049	0.060	0.103	0.117	0.123	0.120
3: Prop. Match.	Coeff	0.223	0.261	0.141	–	–	–
Measurement Error	StdE	0.071	0.094	0.113	–	–	–

Robust standard errors are reported for specification 1; bootstrapped standard errors (based on 500 repetitions) are reported for specifications 2,3.

Altogether, these results indicate that measurement error and endogenous choice bias the estimates of language effects in opposite directions. Our results suggest that the true effect of language on employment probabilities is substantial, and possibly larger than simple OLS estimates suggest. Overall, the results we obtain from the estimator which controls for measurement error suggest that fluency increases the probability that a male individual is employed, given that he looks for a job, by around 26 percentage points. The estimate for females is smaller, and not significant.

4.3. *Earnings*

We now turn to the effect of language on weakly gross earnings. Neither sample provides information on the number of hours worked per week, and we therefore consider only individuals who are working full-time.

In the FWLS, the dependent variable is the natural logarithm of gross (before tax) weekly earnings. The earnings variable in the FNSEM is gross weekly earnings, which is reported in categorical form (16 categories). In both samples there is a considerable percentage of working individuals who do not report their earnings (28% in the FNSEM and 45% in the FWLS).

To check the extent to which attrition is non-random, we compare the means of the language variables, origin dummies, the educational variables and other individual characteristics for individuals who do, and who do not report earnings. Results are presented in Table A3. We also report the t-statistics for testing whether the means of the variables are significantly different. In some cases, we reject the null hypothesis of equal means, but there seems to be no systematic pattern of attrition across the two data sets.

Our final sample sizes for the earnings analysis are 254 individuals for the FWLS data, and 920 individuals for the FNSEM data. Results of straightforward log wage regressions are presented in Table 7, where we use the least squares estimator for the FWLS, and the least squares estimator at the midpoints for the FNSEM.[11]

As regressors, we include the same set of variables as in the employment regressions. Coefficient estimates on most variables are roughly similar for the two data sets. Males have a significant earnings advantage, compared to females. Having a degree more than doubles earnings, compared to holding no qualification. O-levels (or equivalent) alone increase earnings by about 17 (FWLS) or 24% (FNSEM).[12]

In the last column, we again use the more detailed educational information in the FNSEM, and decompose educational attainments into overseas and UK qualifications. We find that the coefficients on UK qualifications are larger than overseas ones, but the joint null hypothesis that degrees acquired abroad have a significantly different effect on earnings from degrees acquired in the UK is

[11] We have also estimated grouped regression models for the FNSEM (where the boundaries are transformed by taking logs). Results are almost identical.

[12] We compute here and in the following percentage differences in earnings as $(e^{\hat{\beta}} - 1) \times 100$, where $\hat{\beta}$ is the estimated parameter on the variable to which the discussion refers.

Table 7

Earnings Regressions

	FWLS						FNSEM			
	1		2		3		4		5	
							All Qualifications		UK/nonUK Q	
Variable	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE	Coeff	StdE
Cons	3.551**	0.411	3.577**	0.412	3.546**	0.413	3.843**	0.243	3.809**	0.249
male	0.238**	0.072	0.251**	0.071	0.238**	0.072	0.107**	0.039	0.115**	0.039
married	−0.010	0.088	−0.008	0.089	−0.008	0.089	0.176**	0.051	0.160**	0.051
degree	0.786**	0.104	0.788**	0.106	0.781**	0.106	0.671**	0.048	–	–
Alev	0.206*	0.090	0.202*	0.093	0.201*	0.093	0.384**	0.051	–	–
OlevCSE	0.169	0.091	0.172	0.091	0.166	0.092	0.156**	0.043	–	–
Edegree	–	–	–	–	–	–	–	–	0.607**	0.056
EAlev	–	–	–	–	–	–	–	–	0.351**	0.054
EOlevCSE	–	–	–	–	–	–	–	–	0.120*	0.050
Fdegree	–	–	–	–	–	–	–	–	0.504**	0.066
FAlev	–	–	–	–	–	–	–	–	0.132	0.078
FOlevCSE	–	–	–	–	–	–	–	–	0.094	0.050
age	0.038	0.023	0.036	0.023	0.038	0.023	0.019	0.012	0.021	0.013
agesq/100	−0.045	0.029	−0.042	0.029	−0.044	0.029	−0.022	0.015	−0.025	0.015
yearst	0.026	0.015	0.030*	0.014	0.027	0.015	0.033**	0.007	0.032**	0.007
$yearst^2/100$	−0.050	0.035	−0.035	0.035	−0.050	0.036	−0.051**	0.019	0.050*	0.020
carib	0.302*	0.132	0.327*	0.130	0.301*	0.132	0.279**	0.076	0.301**	0.077
afroas	0.081	0.125	0.109	0.123	0.083	0.125	0.224**	0.068	0.259**	0.068
indian	0.311**	0.113	0.329**	0.112	0.310**	0.113	0.157*	0.069	0.206**	0.069
pakista	0.239*	0.118	0.251*	0.118	0.239*	0.119	0.025	0.072	0.066	0.073
chinese	–	–	–	–	–	–	0.408**	0.083	0.416**	0.085
speak	0.204	0.115	–	–	0.171	0.161	0.180**	0.055	0.192**	0.055
write	–		0.149	0.103	0.040	0.145	–	–	–	–
No. of Obs.	254		254		254		920		920	

Base Category: no educational qualification, Bangladeshi. Robust standard errors are reported. *Significant at 5% level. **Significant at 1% level.

rejected at the 5% level. Coefficients are only significantly different for A levels or equivalent degrees.

The coefficients on the ethnicity dummies indicate significant wage differences between ethnic groups. As in the language and employment equations, Bangladeshis are the most disadvantaged group. Conditional on education, age and years of residence, their wages are 66% lower than those of the most successful group, the Chinese (FNSEM). In both data sets the earnings of Caribbeans are about 35% higher than Bangladeshis.

We find large and significant coefficients for the English fluency variables. The point estimates in the FNSEM and FWLS are quite similar and indicate that English language proficiency is associated with about 21 (FNSEM) or 23% (FWLS) higher wages. Again, we use writing proficiency as an additional indicator for language proficiency (see columns 2 and 3). Interestingly, and different from the employment equation, fluency seems to be more important for wages than literacy.

4.4. *Earnings, Endogenous Choice and Measurement Error*

Besides measurement error and endogenous choice of acquiring language proficiency, an additional difficulty with investigating earnings is non-random selection into the workforce. Non-participation is large among minority immigrants, in particular among females. It is likely that participation is selective and correlated with the choice to acquiring language proficiency, thus biasing parameter estimates.

The conventional way of addressing non-random selection is to model the selection process and the earnings equation simultaneously. A simple estimator is a two step estimator which conditions earnings on the (generalised) residual from the first step auxiliary participation equation. To implement this approach requires identifying assumptions. We experimented with a number of possible exclusion restrictions. We are not confident about the validity of most exclusion restrictions that are feasible given the information in our data.[13]

We therefore refrain from estimating a joint model. To the extent that the participation choice is due to observables, our matching approach takes care of this problem. For any remaining selection, our strategy is to interpret the coefficients on the language variable as bounds, which is possible under some plausible assumptions. As we have seen in the last Section, language has a positive effect on employment probabilities, and simple regressions show that it has also a positive effect on participation. If we are willing to assume that unobservables, which affect the participation probability, are positively correlated with unobservables which affect earnings, then the estimate of the language coefficient in an earnings regression on participants only is downward biased, compared to the hypothetical coefficient for the overall population. The intuition is simple: those individuals, who are not proficient in the English language but participate nevertheless, must be drawn from the upper part of the ability distribution to compensate for their language deficiencies, thus inducing a downward bias in the estimated language coefficient.[14] Accordingly, we can interpret the coefficient estimates we obtain on the sample of participants as lower bounds of the effect of language on earnings.

In Table 8, we report results for the pooled sample and for males and females separately. Splitting the sample into males and females leads to very small sample sizes, in particular for the FWLS, and most of our estimates are quite imprecise. We should therefore interpret results with care.

[13] For females, we considered to use variation in religious beliefs (conditional on origin) as an instrument for participation. The idea is that some religions may impose a strict role behaviour on females more than others. Religion may thus explain variation in participation. The FNSEM data distinguish between Sikh, Hindu, Muslim, Christian and no religion. These variables are jointly significant in an auxiliary first step participation regression. The generalised residual was not significant in the earnings regression, and hardly changed the language coefficient.

[14] More formally, suppose that the latent participation index p^* is linear in l, with $p_i^* = \alpha_0 + \alpha l_i + u_i$, and that the individual participates if $p_i^* > 0$. Suppose that the outcome equation is given by $y_i = \gamma_0 + \gamma l_i + v_i$, and assume that u_i and v_i are jointly normally distributed, with variances 1 and σ_v^2 and correlation coefficient ρ. Then selection could be accounted for by adding the generalised residual $\mathrm{E}(v_i|p_i^* > 0) = \lambda(c_i)$ to the estimation equation, where $\lambda(c_i) = \phi(c_i)/\Phi(c_i)$, with ϕ and Φ being the density and distribution function of the standard normal, and $c_i = \alpha_0 + \alpha l_i$. We obtain the estimation equation $y_i = \gamma_0 + \gamma l_i + \sigma_v \rho \lambda(c_i) + e_i$. Omission of $\lambda(c_i)$ results in a biased estimate for γ. The expectation of the error term when omitting λ, conditional on l_i, is $\rho\sigma_v \mathrm{E}[\lambda(c_i)|l_i]$. Since λ decreases in c_i, the bias is downward for $\rho > 0$ and $\alpha > 0$.

Table 8

Earnings and Language

Estimation Specification		FNSEM			FWLS		
		All	Males	Females	All	Males	Females
1: OLS	Coeff	0.180	0.121	0.354	0.204	0.173	0.167
	StdE	0.055	0.063	0.120	0.115	0.180	0.121
2: Prop. Match.	Coeff	0.281	0.238	0.463	0.101	–	–
	StdE	0.108	0.103	0.186	0.174	–	–
3: Prop. Match.	Coeff	0.356	0.460	0.844	–	–	–
Measurement Error	StdE	0.324	0.272	0.844	–	–	–

Robust standard errors are reported for specifications 1; bootstrapped standard errors (based on 500 repetitions) are reported for specifications 2,3.

In the first row, we report the least squares results. While for the FWLS, coefficients for males and females are quite similar, the language coefficient using the FNSEM data is much larger for females than for males. Coefficient estimates for the FWLS are however not significant, with large standard errors for the separated samples.

In the second row, we report results from the propensity score estimator. Coefficients for both males and females are larger relative to the simple OLS estimator. This seems to be contrary to what endogenous choice of language acquisition would predict. However, as we discussed above, non-random participation may lead to downward biased estimates of language coefficients. The matching estimator corrects for participation selection, as long as it is on observables, and may therefore reduce the downward bias due to selective participation. Sample sizes for the FWLS data when we distinguish between males and females became too small for this estimator, and we only report results for the FNSEM.

In row 3, we implement our estimator which accounts for measurement error in addition. For females, the coefficient estimate becomes very large, and is estimated with very low precision. For males, coefficient estimates increase by factor 2, but the coefficient is not significant at the 5% level. Sample sizes are too small to draw robust conclusions from this evidence. We may however interpret the increase in coefficients when correcting for measurement error as evidence that measurement error leads to downward biased estimates also here.

5. Summary and Discussion

Based on two recent UK surveys, we analyse the determinants of English language fluency for ethnic minority immigrants in the UK and the effect of language on labour market outcomes. We also investigate the effect of other characteristics on language acquisition, and employment and earnings.

We find that in simple regressions, language proficiency is associated with higher employment probabilities and with higher earnings. Language effects may be under or overestimated, due to endogenous choice of learning the language,

and measurement error. We address both these issues. We use a matching estimator to address the endogenous choice of language acquisition. We combine our matching estimator with an IV type estimator to eliminate the downward bias due to measurement error, using information about the interview language for identification. Our results indicate that the bias induced by the two problems points in opposite directions, and that the effect of language on outcomes is larger than suggested by simple regression estimators. While OLS estimates indicate that language fluency increases employment probabilities by 17 percentage points, estimates that address both selection and measurement error suggest an increase by about 22 percentage points. Our analysis on earnings is less conclusive. OLS estimates suggest an earnings advantage of those who are proficient in English of about 18–20%. Estimates based on the estimator that addresses both endogenous selection and measurement error are insignificant.

The validity of our matching approach depends on our beliefs about whether the set of matching variables eliminates the problem of endogenous selection. The set of conditioning variables available to us includes indicators that are likely to be correlated with unobserved ability that sorts individuals into groups of those who do and who do not acquire the host country language, like education, ability tests and partner information. However, if these variables do not fully account for unobserved factors that select individuals into the group of those who are proficient and non-proficient in the English language, language effects may still be upward biased.

Addressing the problem of endogenous language choice is difficult. Ideally, we would like to observe immigrants who have had different access to language facilities and where the assignment to facilities is exogenous. One mechanism that could generate this are settlement policies that allocate immigrants to different communities upon arrival. Schemes like this were in place in different countries. Future research could use these assignment mechanisms to address the problem of endogenous language choice.

As we discussed in the Introduction, the data we use in this analysis do not cover the entire immigrant population in the UK but only those immigrants who belong to ethnic minority communities. According to the Labour Force Survey (2000), immigrants from ethnic minority groups constitute only 49% of the total immigrant population in the UK. Hence, our analysis covers only half of the immigrant population. As shown by other research on UK immigrants (Chiswick, 1980; Bell, 1997; Wheatley Price, 2001; Dustmann *et al.*, 2002), the assimilation patterns of ethnic minority immigrants and white immigrants differ quite substantially. It is likely that effects of language proficiency on economic outcomes are also different for these groups. More comprehensive surveys are needed to allow investigating language effect for the entire immigrant population in the UK.

University College, London

Date of receipt of first submission: March 2000
Date of receipt of final typescript: May 2002

Appendix

Table A1

Language Information

	All groups	Caribbean	Indian	Afroasian	Pakistani	Bangladeshi	Chinese
Speaking, FWLS							
Very well	37.81	54.55	50.44	64.77	38.16	25.93	–
Quite well	23.12	13.64	27.43	27.27	26.05	18.46	–
Not well	20.12	18.18	18.14	5.68	21.32	22.82	–
Hardly	11.69	13.64	3.54	2.27	10	18.46	–
Not at all	7.26	–	0.44	–	4.47	14.32	–
Reading, FWLS							
Very well	34.64	40.91	48.67	61.36	33.16	24.07	–
Quite well	21.12	18.18	23.89	26.14	21.58	18.67	–
Not well	15.86	22.73	14.16	7.95	17.11	16.8	–
Hardly	13.19	9.09	7.96	1.14	14.47	17.01	–
Not at all	15.19	9.09	5.31	3.41	13.68	23.44	–
Writing, FWLS							
Very well	32.39	40.91	45.13	56.82	29.47	23.86	–
Quite well	19.2	18.18	21.68	23.86	20.79	15.98	–
Not well	16.61	22.73	15.49	13.64	18.16	16.18	–
Hardly	21.77	4.55	11.06	2.27	13.68	15.15	–
Not at all	19.03	13.64	6.44	3.41	17.89	28.84	–
Speaking, FNSEM							
Fluent	48.73	86.95	39.98	65.63	25.56	25.97	56.59
Fairly	20.4	9.62	24.37	19.2	25.56	23.02	12.64
Slightly	21.2	–	25.84	11.76	32	34.25	18.13
Not at all	9.67	–	9.81	3.41	16.88	16.76	12.64

Table A2

Age and Time Patterns

	Age Migration		Year		Age	
Ethnicity	FWLS	FNSEM	FWLS	FNSEM	FWLS	FNSEM
Carib	19.460	20.379	1964.2	1963.435	49.755	50.927
StdD	8.840	10.186	7.968	7.772	12.199	13.933
Indian	18.971	23.892	1973.2	1972.431	40.299	45.145
StdD	10.235	13.535	9.750	9.515	11.004	14.384
AfroAsia	17.813	20.662	1974.1	1973.828	37.976	40.735
StdD	10.320	12.702	7.035	7.221	10.393	13.026
Pakista	18.167	20.424	1976.4	1974.207	35.870	39.672
StdD	9.374	11.275	9.634	9.677	11.953	13.738
Bangla	18.676	20.579	1979.6	1977.695	33.266	36.645
StdD	9.374	10.545	9.046	9.637	13.985	14.156
Chinese	–	22.088	–	1976.35	–	39.641
StdD	–	11.860	–	8.670	–	12.532
All	18.663	21.418	1974.7	72.414	38.308	42.707
StdD	10.084	11.918	1.402	9.944	13.587	14.572

Table A3

Attrition

Variable	FWLS Report Mean	FWLS Missing Mean	FWLS Diff. t-value	FNSEM Report Mean	FNSEM Missing Mean	FNSEM Diff. t-value
sex	0.618	0.730	2.59	0.695	0.682	0.46
married	0.767	0.802	0.91	0.840	0.861	1.04
nchild	1.480	1.995	3.53	1.484	1.503	0.20
degree	0.153	0.110	1.36	0.2	0.230	1.22
Alevtea	0.212	0.144	1.92	0.163	0.167	0.18
OlevCSE	0.208	0.278	1.74	0.269	0.242	1.06
age	37.704	38.274	0.59	39.358	40.726	2.32
yearstay	22.303	22.783	0.54	22.021	22.542	1.00
black	0.204	0.182	0.59	0.229	0.101	6.27
afroas	0.145	0.129	0.49	0.244	0.227	0.67
indian	0.338	0.264	1.73	0.225	0.382	5.70
pakista	0.173	0.278	2.69	0.140	0.181	1.86
Chinese	–	–	–	0.078	0.029	4.05
speak	0.877	0.865	0.40	0.877	0.876	0.03
write	0.850	0.793	1.58	–	–	–
No of Obs.	254	208		413	920	

Note: t-statistics computed as $(m_1 - m_2)/\sqrt{se_1^2 + se_2^2}$, where m_i, se_i are means and standard errors of the two sample values, respectively.

References

Bell, B. D. (1997). 'The performance of immigrants in the United Kingdom: evidence from the GHS', Economic Journal, vol. 107, pp. 333–44.

Berman, E., Lang, K. and Sriniver, E. (2000). 'Language-skill complementarity: returns to immigrant language acquisition', mimeo, Boston University.

Blackaby, D. H., Clark, K., Leslie, D. G. and Murphy, P. D. (1994). 'Black-white male earnings and employment prospects in the 1970s and 1980s evidence for Britain', *Economics Letters*, vol. 46, pp. 273–9.

Blackaby, D. H., Drinkwater, S., Leslie, D. G. and Murphy, P. D. (1997). 'A picture of male and female unemployment among Britain's ethnic minorities', *Scottish Journal of Political Economy*, vol. 44(2), pp. 182–97.

Borjas, G. J. (1985), 'Assimilation, changes in cohort quality, and the earnings of immigrants', *Journal of Labour Economics*, vol. 3, pp. 463–89.

Borjas, G. J. (1987). 'Self-selection and the earnings of immigrants', *American Economic Review*, vol. 77, pp. 531–53.

Borjas, G. J. (1994). 'The economics of immigration', *Journal of Economic Literature*, vol. 32, pp. 1667–717.

Chiswick, B. R. (1978). 'The effect of Americanization on the earnings of foreign-born men', *Journal of Political Economy*, vol. 86, pp. 897–921.

Chiswick, B. R. (1980). 'The earnings of white and coloured male immigrants in Britain', *Economica*, vol. 47, pp. 81–7.

Chiswick, B. R. (1991). 'Speaking, reading, and earnings among low-skilled immigrants', *Journal of Labor Economics*, vol. 9, pp. 149–70.

Chiswick, B. R. (1994). 'Language choice among immigrants in a multilingual destination', *Journal of Population Economics*, vol. 7, pp. 119–31.

Chiswick, B. R. and Miller, P. W. (1995). 'The endogeneity between language and earnings: international analyses', *Journal of Labor Economics*, vol. 13, pp. 246–88.

Chiswick, B. R., Cohen, Y. and Zach, T. (1997). 'The labor market status of immigrants: effects of the unemployment rate at arrival and duration of residence', *Industrial and Labor Relations Review*, vol. 50(2), pp. 289–303.

Clark, K. and Drinkwater, S. (2000). 'Pushed out or pulled in? Self-employment among ethnic minorities in England and Wales', *Labour Economics*, vol. 7, pp. 603–28.
Dearden, L. (1999). 'Qualifications and earnings in Britain: how reliable are conventional OLS estimates of the returns to education?', IFS working paper no. W99/7.
Dearden, L., Ferri, J. and Meghir, C. (2000) 'The effect of school quality on educational attainment and wages', forthcoming *Review of Economics and Statistics*.
Dustmann, C. (1994). 'Speaking fluency, writing fluency and earnings of migrants', *Journal of Population Economics*, vol. 7, pp. 133–56.
Dustmann, C. (1999). 'Temporary migration, human capital, and language fluency of migrants', *Scandinavian Journal of Economics*, vol. 101, pp. 297–314.
Dustmann, C. and van Soest, A. (2001). 'Language fluency and earnings estimation with misclassified language indicators', *Review of Economic and Statistics*, vol. 83, pp. 663–74.
Dustmann, C., Fabbri, F., Preston, I. and Wadsworth, J. (2002). 'The performance of immigrants in the UK', report for the Home Office at www.homeoffice.gov.uk/rds/pdfs2/rdsolr0505.pdf.
Hatton, T. J. and Wheatley Price, S. (1999). 'Migration, migrants and policy in the United Kingdom', IZA discussion paper no. 81.
Heckman, J. (1997). 'Instrumental variables', *Journal of Human Resources*, vol. 32, pp. 441–62.
Heckman, J., Ichimura, H. and Todd, P. (1998). 'Matching as an econometric evaluation estimator', *Review of Economic Studies*, vol. 65, pp. 261–94.
Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1997). 'Matching as an econometric evaluation estimator: evidence from evaluating a job training programme', *Review of Economic Studies*, vol. 64, pp. 605–54.
Modood, T. and Berthoud, R. (1997). *Ethnic Minorities in Britain*, London: Policy Studies Institute.
Nickell, S. J. (1980). 'A picture of male unemployment in Britain', Economic Journal, vol. 90, pp. 776–94.
Rivera-Batiz, F. L. (1990). 'English language proficiency and the economic progress of immigrants', *Economic Letters*, vol. 34, pp. 295–300.
Rosenbaum, P. and Rubin, D. B. (1983). 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, vol. 70, pp. 41–55.
Shields, M. A. and Wheatley Price, S. (1998). 'The earnings of male immigrants in England: evidence from the quarterly LFS', *Applied Economics*, vol. 30, pp. 1157–68.
Shields, M. A. and Wheatley Price, S. (2002). 'The English language fluency and occupational success of ethnic minority immigrant men living in English metropolitan areas', *Journal of Population Economics*, vol. 15, pp. 137–6.
Smith, P. and Prior, G. (1996). *The Fourth National Survey of Ethnic Minorities: Technical Report*, London: Social and Community Planning Research.
Stewart, M. B. (1983). 'On least squares estimation when the dependent variable is grouped', *Review of Economic Studies*, vol. 50, pp. 737–53.
Wheatley Price, S. (2001). 'The employment adjustment of male immigrants to the English labour market', *Journal of Population Economics*, vol. 14, pp. 193–220.