The Relationship between Reading Proficiency and Earnings: A Study among UK Immigrants and Ethnic Minorities

Student number: 1403028
Supervisor: Christian Soegaard
Word count: 4936

Research in Applied Economics
Department of Economics
University of Warwick

2017/18

Abstract

This paper utilizes contemporary cross-sectional data to analyse the causal effect of achieving complete reading fluency on earnings for UK immigrants and ethnic minorities, for whom English is not their first language. In addition to an initial OLS specification, two further econometric methods are used in order to correct for endogeneity and self-selection problems that arise from incorporating language variables in earnings regressions. Results suggest that achieving complete reading fluency provides a significant wage premium and that reading deficiencies act as a prominent earnings constraint both within and between industries.

1. Acknowledgements

The author would like to sincerely thank Dr. Christian Soegaard for the advice and guidance provided with regards to data, econometric analysis and presentation of findings. Guidance provided by Dr. Claire Crawford is greatly appreciated with regards to the use of propensity score matching.

Contents

2. Introduction

The number of non-UK born residents increased by 0.6 million between 2015 and 2016, constituting 14% of the total UK population in 2016 (Office for National Statistics, 2017). Moreover, for those individuals for which English is not their main language, the employment rate was 17.1% higher for those categorised to be “proficient” compared to “non-proficient” (Office for National Statistics, 2014). These descriptive statistics highlight the need to examine the economic relationship between language and labour market outcomes, as immigrants continue to constitute an even higher proportion of the UK labour force. Furthermore, in light of the plan to devolve adult education budgets to Mayoral Combined Authorities in 2019 (Gov.uk, 2017), it is vital that research is available to inform policy makers of the importance of facilitating language learning in highly concentrated immigrant areas, thus ensuring that the UK labour market continues to be flexible in future years.

Moreover, the UK government has indicated that it will adopt a more stringent immigration policy in a post-Brexit Britain (Ft.com, 2017). Currently, immigrants are required to have a B1 level (intermediate user) of reading, writing, listening and speaking skills in English to obtain a Tier 2 general visa in the UK. Examining the relationship between reading fluency and earnings thus provides important insight for the future screening process for prospective workers. Results indicate the necessity to increase English language requirements due to the economy-wide benefits a more linguistically equipped labour force would provide.

The objective of this paper is to ascertain whether achieving complete reading fluency has a significant positive impact on monthly wages for UK immigrants and ethnic minorities, for whom English is not their first language. A secondary objective is to discover if reading deficiencies hamper earnings prospects both within and between high and low skilled industries. The data that will be used for this analysis is 2016 cross-sectional data from wave 6 of *Understanding Society* (UK household longitudinal study).

Ordinary Least Squares (OLS) estimation suggests that an increase in reading proficiency from non-fluent to fluent leads to a 30.2% increase in monthly earnings. Instrumental Variable (IV) and Propensity Score Matching (PSM) techniques have been applied to correct for endogeneity and self-selection issues. IV estimation confirms a significant positive impact, suggesting that complete reading fluency increases monthly earnings by 41.3%. Finally, PSM analysis concludes that reading deficiencies constrain an individuals earnings potential, as those who are not fluent on average earn £204 less per month than those who are. Reading deficiencies have been shown to constrain earnings potential in both high and low skilled industries by £154 and £231 respectively.

This paper contemporises the UK literature available on the topic as it is first paper focusing on the subject matter since 2003. Moreover, it adds to the literature available by extending the analysis to immigrants across a wider range of nationalities than immigrants solely from Asian and Caribbean descent, incorporating individuals born in the EU and The Americas. Finally, to the best of the author's knowledge, this will be the first paper using UK data to focus solely on the effect of English reading fluency upon earnings.

3. Literature review

The existing literature on the effect of English language proficiency on earnings stems from the work conducted by Chiswick (1978), where the effect of length of time spent in the US on the earnings of non-US born immigrants is examined. This paper was the basis for further research focusing more specifically on the effect of language on earnings, which began solely for Hispanic men by Reimers (1983) and Grenier (1984). Moreover, Mcmanus, Gould and Welch (1983) highlighted the importance of communication as a form of human capital, both for factors of production to work efficiently and for the provision of a high-quality service to customers, thus improving producer earnings. Tainer (1988) expanded on the work conducted in the early 1980s by affirming a significant positive effect of improved English language proficiency on earnings for all ethnic groups within the US, due to higher proficiency boosting productivity levels. A unitary increase in speaking proficiency, measured on a scale from one to four, caused a 17.4%, 17.2% and 12.7% increase in annual earnings for Hispanics, Asians and Europeans respectively. Despite a more inclusive range of ethnic groups in this study, the subjectivity of the self-reported speaking variable, used to measure proficiency, is a cause of measurement error (Dustmann and Van Soest, 2002).

Rivera-Batiz (1990) corrected this issue by using an objective measurement of reading ability using test scores from the National Assessment of Educational Progress. His results echo the importance of reading proficiency on earnings, suggesting that reading deficiencies were a major factor in constraining labour mobility between low and high paid industries. Chiswick's work (1991) was superior to earlier studies due to the inclusion and comparison of reading and speaking variables in the same regression analysis and found that reading "well" or "very well" led to an increase in weekly earnings by 31%, compared to a statistically insignificant coefficient for the speaking variable. These two papers highlighted the importance of reading proficiency over and above speaking fluency alone, thus providing a key motivation to explore this particular aspect of language proficiency in this paper.

3.1. Statistical shortcomings

Building on studies in the 1980s and 90s, the statistical problems of endogeneity and measurement error were addressed to add further econometric rigour to previous findings. It is plausible that those with greater intellectual ability choose to be more proficient in English and earn more (Borjas, 1994), therefore creating an endogeneity bias when using an OLS specification. To address this issue Chiswick and Miller (cited in Borjas, 1994) used number of children and fraction of people in the state that speak the same language as instruments. Moreover, Dustmann and Fabbri (2003) used a combination between matching and IV estimation to deal with this problem. Dustmann and Van Soest (2002) used instruments such as lags and leads of self-reported language proficiency to address the measurement error concern, becoming the first authors to use panel data methodology to address this problem. In this analysis, IV and PSM techniques are employed to add robustness to an initial OLS specification.

3.2. Contemporary findings

The impact of language proficiency upon earnings is not confined to the UK. Contemporary literature in the 21st century built on the aforementioned ideas and methodologies, using datasets from different countries and geographic regions. Findings in Australia (Guven and Islam, 2015), the US (Hamilton, 2014) (Hwang, Xi and Cao, 2010), Spain (Di Paolo and Raymond, 2012) (Budria and Swedberg, 2015) and Europe (Ginsburgh and Prieto-Rodriguez, 2013) all focus on the importance of language capital as a determinant of immigrant and native earnings. However, a lack of suitable data containing language and earnings variables has constrained research in the UK to be contemporised to the level within Europe and the United States. Dustmann and Fabbri (2003) found that proficiency in English increased earnings by 22%, as an average across two UK datasets. This research is limited to immigrants from Asian and Caribbean descent. The inclusion of immigrants with European origins signifies that this paper is more indicative of the current immigrant demographic in the UK, given the increase in EU net migration by 49% from 2007 to 2016 (Office for National Statistics, 2018).

4. Hypotheses

The acquisition of host-country human capital is vital for the economic assimilation of the migrant population (Borjas, 1994). Reading proficiency is a key element of this, and can affect earnings via the following channels:

1. Language proficiency boosts the productivity of immigrants within the labour force (Tainer, 1988). Reading proficiency, particularly within highly technical occupations, will facilitate the analytical capabilities of the immigrant workforce, thus enhancing intra-industry (within the same industry) promotion prospects and earnings potential.
2. Command of the host country language improves job search prospects, as migrants will be naturally more aware of the range of opportunities available to them (Dustmann, 1994). Proficiency in reading will allow immigrants to take advantage of various forms of job advertising to ascend in their industry of expertise, allowing them to earn a higher wage.
3. Language deficiency undoubtedly constrains upward labour mobility into higher-paid industries, hence earnings prospects (Rivera-Batiz, 1990). A lack of reading proficiency may be a significant barrier reducing the probability of success in literacy-heavy recruitment processes, thus reducing inter-industry (between industries) mobility.

All three channels indicate that improved reading proficiency will increase earnings. However, channels 1 and 2 suggest that reading deficiencies hamper intra-industry mobility, and the third indicates a reduction in inter-industry mobility. Thus, the aforementioned theoretical frameworks give rise to the following hypotheses:

Hypothesis 1: Achieving complete reading fluency causes a significant positive impact on monthly wages.

Hypothesis 2: Reading deficiencies constrain earnings potential both within and between industries.

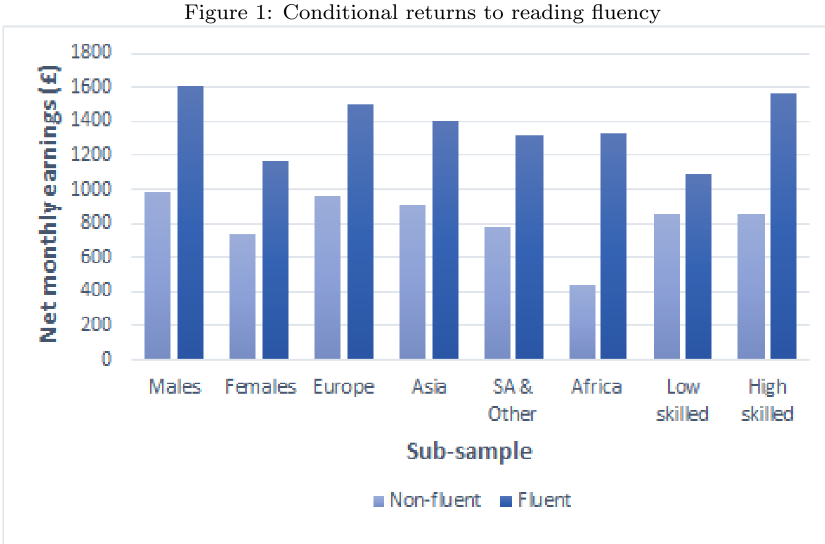
5. Dataset

The subsequent analysis is based on cross-sectional data from wave 6 of *Understanding Society* (the UK Household Longitudinal Study). It is based on the adult interview section of the survey questionnaire. This wave includes a comprehensive Immigrant and Ethnic Minority Boost sample (IEMB), consisting of 4,656 respondents, who have answered specific questions relating to language proficiency and earnings. The results are focused purely on this boost sample, hence only reflect the causal impact of proficiency on earnings for those immigrants and ethnic minorities living in high immigrant concentration areas. The final sample size available for econometric analysis is 747 respondents, due to a large proportion of immigrants and ethnic minorities being unwilling or unable to provide information on language and earnings. The dataset is highly conducive to the usage of a matching technique as it can reduce the bias level of results (Rosenbaum, 2005). This is because the immigrants included in this survey are likely to share many personal demographic, family and labour market characteristics which natives do not, such as country of origin, ethnic group and time spent in the UK.

Figure 7 in the appendix shows a representation of the sample in question. 28% of the sample are from Europe and Africa, a sub-sample never previously analysed in the UK literature. Moreover, this dataset includes immigrants from Canada, US, New Zealand and Australia, forming a part of the "South America and Other" category, which constitutes 36% of the total sample. 7% of the sample used are ethnic minorities born in Britain, but for whom English is not their first language. These respondents are included to avoid reduced sample size.

Descriptive statistics show that those categorised as non-fluent earn a mean net monthly wage of £859, compared to those who are fluent earning £1387 on average. The difference in means is significant at the 1% level (Appendix, Figure 8). This is in line with the prediction of hypothesis 1, suggesting that immigrants and ethnic minorities who have obtained complete fluency earn higher wages. This difference in means remains significant across gender and continent sub-samples at the 5% level, as shown in figure 1 below.

Figure 1: Conditional returns to reading fluency



For high and low skill occupations respectively, those who are not fluent earn on average £705 and £235 less per month than those who are. This result is significant at the 1% and 10% level respectively, suggesting that reading deficiency does hamper earnings potential within industries, as predicted by hypothesis 2.

6. Methodology and Results

This section of the paper will be structured as follows. Firstly, Ordinary Least Squares (OLS) estimation will be used as a baseline model. Secondly, this result will be complemented using Instrumental Variable (IV) estimation to correct for endogeneity bias present between language fluency and earnings (Chiswick and Miller, 1995). These two econometric methods will substantiate the claim made in hypothesis 1. Finally, the use of Propensity Score Matching (PSM) will be used as a further robustness check and as a rigorous method to legitimize the claims of hypothesis 2.

The independent variable of interest, *f_engread*, a dichotomous variable representing complete reading fluency, takes the value 1 if the respondent states that they have no difficulty in reading formal documents or letters in English. This is what will be and has been referred to as fluent in terms of reading proficiency. The dependent variable is net monthly earnings, transformed into logarithmic form in order to normalise the distribution of the earnings variable. The regressand will be referred to as earnings in the subsequent text.

6.1. Ordinary Least Squares (OLS)

As a basis for further analysis, an OLS specification was built via the inclusion of groups of exogenous control variables. Specifications 2, 3 and 4 build on the first, which only includes the independent variable of interest, by including additional groups of controls for personal demographic characteristics, family demographic characteristics and labour market traits to reach a final specification as seen in column 4 below. A complete list of coefficient estimates, and a list of control variables by group are detailed in figures 9,10 and 11 in the appendix. Coefficient estimates are reported at a constant sample size of 554 to avoid bias due to missing values of the exogenous controls. Finally, a model with a very similar econometric specification to that of the latest paper in the UK was estimated. This result is reported in the final column to serve as a robustness check.

Figure 2: OLS earnings regressions

VARIABLE	OLS (1) Log of income	OLS (2) Log of income	OLS (3) Log of income	OLS (4) Log of income	OLS Dustmann & Fabri (2003)
f_engread	0.369***	0.481***	0.452***	0.264***	0.276***
	(0.0619)	(0.0614)	(0.0635)	(0.0741)	(0.0692)
+ controls	-	-	-	-	-
Constant	6.675***	5.818***	4.006***	4.423***	4.538***
	(0.0535)	(0.187)	(0.335)	(0.346)	(0.351)
Observations	554	554	554	554	554
R-squared	0.037	0.139	0.233	0.376	0.254

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficient estimates for *f_engread* remain significant at the 1% level across all models. Drivers for the reduction in coefficient value from specification 3 to 4 include large positive coefficients for occupation type and residing in London, which have been shown to have a larger semi-elasticity with respect to wages than family or personal demographic characteristics. The key finding is reported in column 4, demonstrating that a movement from non-fluent to fluent in terms of English reading proficiency causes a 30.2% increase in earnings. This corroborates the finding using alternative control variables as used by Dustmann and Fabbri, showing a 31.8% increase in earnings. A final robustness check included re-estimating the model with the removal of the top 1% of earners in the dataset, of which coefficient estimates remained significant at the 1% level. These results affirm the claim made by hypothesis 1.

6.2. Instrumental Variable (IV) Estimation

There is a strong possibility that those respondents who are naturally more capable are more likely to obtain language fluency and earn more (Borjas, 1994). Due to the inherent endogeneity issue and the lack of a test score variable to control for ability, binary instruments for whether the respondent spoke English prior to arrival to the UK and conversational speaking fluency were used to instrument for English reading proficiency. It is plausible that both instruments will be relevant in determining reading fluency and are unlikely to directly impact earnings, due to the instruments only capturing partial fluency and whether English was spoken prior to the assimilation process. The results are detailed below.

Figure 3: IV estimation output

VARIABLES	IV (4) Log of earnings
f_engread	0.347*
	(0.194)
Constant	4.483***
	(0.378)
Observations	554
R-squared	0.356

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

IV estimation suggests that a movement from non-fluent to fluent in terms of reading proficiency causes a 41.3% increase in earnings. Tests for instrumental relevance and exogeneity were passed, as detailed in the appendix. Despite this statistical backing, due to the causal link identified between complete speaking fluency and earnings, as shown by Tainer (1988), the instruments are not perfect, thus IV estimates must be interpreted with caution. Nevertheless, the result further substantiates the claim of hypothesis 1, as the coefficient estimate remains significant at the 10% level under an alternative econometric specification.

Both coefficient estimates for OLS and IV estimations are higher than the majority of literature available, and this can be attributed to a difference in categorisation of the variable for reading fluency in this analysis and a reduction of measurement error. Previous authors have created dichotomous variables placing those who speak or read English "not very well at all" or "not very well" into a non-proficient category and those who speak or read "very well" or "extremely well" into a proficient one. By answering a "yes" or "no" question regarding difficulty of reading formal documents or letters, this creates a binary variable, clearly distinguishing between those who are fluent and those who are not. Bleakley and Chin (2004) have proven that the acquisition of complete fluency gives rise to higher wage premiums than for partial fluency. This is a reason for higher coefficient estimates in this paper.

Moreover, previous literature uses a measure from 1 to 3 or 4 using questions such as "How well do you speak/read English" to determine fluency. This has been stated to cause measurement error due to a respondent placing themselves in a lower than true proficiency category (Bleakley and Chin, 2004). The usage of a simple "yes" or "no" question reduces the possibility of this downward bias, constituting a further reason for higher coefficient estimates in this paper. Finally, the existence of higher IV estimates relative to OLS is consistent with the literature (Isphording, 2013).

6.3. Propensity Score Matching (PSM)

The use of Propensity Score Matching (PSM) is justified to correct for the problem of *"endogenous choice of language acquisition"* (Dustmann and Fabbri, 2003). This is the idea that respondents, due to their individual differences, will self-select themselves into being proficient in the host-country language. Therefore, if it is the case that self-selection causes a difference in earnings, as opposed to the development of language capital, OLS will not produce an unbiased estimator for the causal effect of reading proficiency on earnings (Dustmann and Fabbri, 2003). The matching approach seeks to resolve this issue by creating a propensity score to match only like-for-like individuals, by taking into account the differences in observed characteristics that produce the self-selection bias. The matching method is also seen to be superior to OLS as it does not require assumptions regarding correct functional form (Heckman et al., 1999). Moreover, hypothesis 2 can be addressed with statistical rigour via the use of PSM.

This semi-parametric technique is underpinned by the following assumptions.

1. *Unconfoundedness/Conditional Independence*: It must be assumed that differences in outcome values for those individuals who have and have not participated in treatment, who have the same values for covariates, is caused by treatment. This implies that enough variables that influence both treatment assignment and the outcome variable must be observed by the researcher. This is the primary assumption for PSM (Caliendo and Kopeinig, 2008).
2. *Common support*: This secondary assumption states that there must exist both treated and untreated individuals who have the same propensity score. Alternatively, this implies that the distributions of treatment and control groups must be sufficiently similar for the matching procedure to prove effective. (Heckman et al., 1999)

The fulfillment of these assumptions will be discussed in the subsequent sensitivity analysis.

To effectively implement the matching strategy, it was deemed necessary to construct a new variable *nonproficient*, taking the value 1 if the individual was non-fluent in reading English, the reverse of the dichotomous variable used in prior OLS and IV analysis. This is due to 86% of the sample being defined as completely fluent. Non-fluent and fluent individuals will be referred to as the treated and control group respectively. A logistic model was constructed to estimate the propensity score, which estimates the probability of being nonproficient in English, using all explanatory variables from the initial OLS analysis and the inclusion of determinants for reading proficiency. It is these variables that take into account the characteristics that may cause self-selection bias, which can be seen in figures 13 and 14 in the appendix.

The procedure was conducted using three separate matching algorithms, namely nearest-neighbour, kernel and radius matching. Nearest-neighbour has the proven benefit of reducing bias, thus providing a higher quality match. The kernel matching weighting strategy has the advantage of reducing variance of matched pairs. Radius matching incorporates many of the benefits of both prior algorithms by only selecting matches within a certain range when good quality matches are not available (Caliendo and Kopeinig, 2008). A common support has been used under all matching algorithms to only include matched pairs where overlap exists between the distributions of the treatment and control groups. Moreover, replacement is used so that the same observation in the control group can be used for multiple matches (Dehejia and Wahba, 2002), which is necessary due to the small sample size in this analysis. In order to present robust standard errors, bootstrapping with 50 repetitions was used to account for the extra variation imposed by the usage of a propensity score and common support. All matching methods deliver an estimate for the *"Average Treatment Effect on the Treated"* (Caliendo and Kopeinig, 2008) (ATT), which is interpreted as a difference in expected earnings between non-fluent and fluent individuals, for those who participated in treatment (Caliendo and Kopeinig, 2008), which in this analysis are those individuals who are non-fluent. The figure below presents findings to capture the benefits of all three methods.

Figure 4: PSM results

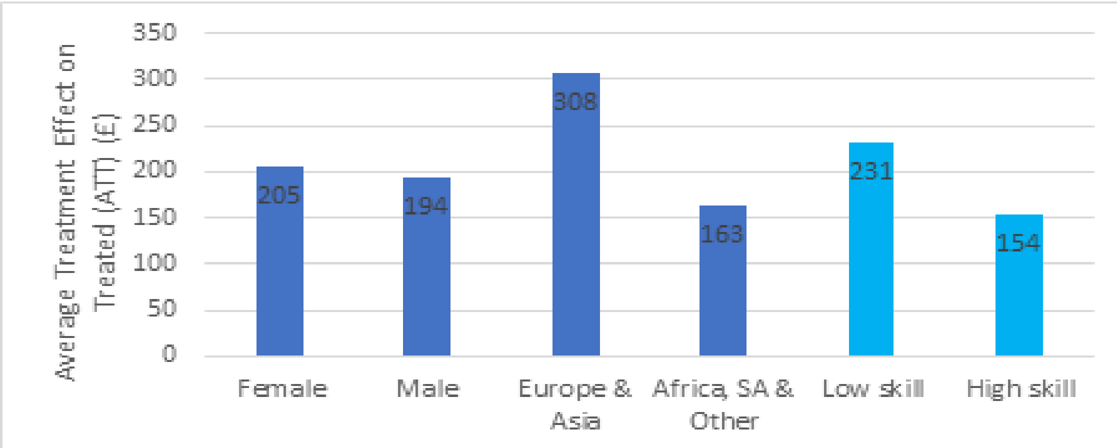
Matching algorithm	**Treated** Log of income	**Control** Log of income	**ATT** Log of income	**T statistic**
Nearest Neighbour	6.67	6.91	-0.25** (0.16)	-2.25
Kernel	6.67	6.89	-0.22** (0.12)	-2.16
Radius	6.67	6.90	-0.23** (0.10)	-2.25

** - significant at the 5% level
Caliper set at 0.05
Bootstrapped standard errors reported in parentheses in difference column

Under nearest neighbour, kernel and radius matching respectively, ATT estimates are -0.25, -0.22 and -0.23 respectively and all results are significant at the 5% level. This indicates that those individuals who are non-fluent in terms of reading proficiency on average earn £214, £194 and £204 less per month than those like-for-like individuals, according to their propensity score, who are fluent in English. As an average across three matching algorithms, this suggests that reading deficiencies constrain earnings potential by £204 per month. This can be extended, assuming a fixed monthly pay structure, to a figure of £2,496 less per annum.

Finally, radius matching was conducted under gender, continent and low and high skilled industry sub-samples, to ascertain whether the initial descriptive statistics affirming hypothesis 2 could be maintained under the use of this econometric technique.

Figure 5: PSM sub-sample analysis



All results remain significant at the 5% level, excluding high-skilled occupations which is significant at the 10% level. The results suggest that reading deficiency constrains earnings prospects within industries as well as between industries, demonstrated by an ATT in high and low skilled occupations of £154 and £231 respectively. These results corroborate the aforementioned descriptive findings, thus confirming the claims of hypothesis 2. This matching approach is not directly comparable to findings in the literature available due to matching either not being used or being used in combination with an IV estimator.

7. Sensitivity Analysis

In addition to bootstrapping and sub-sample analysis, the following checks were undertaken to ensure the robustness of prior matching results.

7.1. Caliper Adjustments

The usage of a caliper signifies that bad quality matches will be excluded from the matching process. A tight caliper has the effect of reducing bias of ATT estimates (Cochran and Rubin, 1973). Due to the difficulty of knowing the correct caliper range prior to matching (Smith and Todd, 2005), calipers were adjusted between 0.15, 0.1 and 0.05 for nearest neighbour and radius matching and a bandwidth between 0.06, 0.05 and 0.03 for kernel matching. No significant changes to coefficients or standard errors were experienced with these adjustments, ensuring low levels of bias in the results.

7.2. Propensity Score Test

A propensity score test was conducted to check the conditional independence and common support assumptions. The outcome of the test provides an indication that the matching procedure has been effective, as it suggests that the covariates used in the logistic regression were of a good quality to successfully predict the propensity scores (Caliendo and Kopeinig, 2008). The results are displayed below.

Figure 6: Propensity score test - Key Statistics

Matching algorithm	Pseudo R^2	$P > chi^2$
Nearest Neighbour	0.089	0.904
Kernel	0.032	1.00
Radius	0.020	1.00

Low Pseudo R^2 and high $p > chi^2$ values both indicate that the observables used to construct the propensity score are good quality, as it has been proven to be difficult to distinguish between the covariates in treatment and control groups. This is a statistical representation of what can be found graphically, by inspecting the degree of overlap between the distributions of propensity scores between control and treatment groups (Lechner, 2008). The distributions pre and post kernel matching are shown in figure 15 in the appendix.

7.3. Rosenbaum Bounds

Sensitivity analysis via the usage of Rosenbaum bounds acts as a final robustness check. This is by ensuring that the existence of unobservables does not affect earnings and non-proficiency in English, thus causing bias to matching estimates. By conducting this check, the author is able to ascertain whether unobservable characteristics not present in the dataset are affecting the self-selection process, thus undermining the conditional independence assumption (Becker and Caliendo, 2007).

Figures 16, 17 and 18 in the appendix confirm that the upper and lower bounds for significance for the Hodges-Lehman point estimate remain significant at the 10% level as gamma (γ) increases from 1 to 1.5. This suggests that as the probability of being assigned into the treatment group increases by 50% due to an unobserved covariate (Rosembaum, 2005), the median differences observed between control and treatment groups continue to be significant for all three matching algorithms. Although this suggests that the results are not completely *"insensitive to bias"* (Rosembaum, 2005), they are maintained to an extent as unobserved variation is increased.

8. Limitations and Extensions

The sample design has limited the scope of analysis to solely respondents in highly concentrated immigrant areas, hence cannot be generalised to the whole immigrant and ethnic minority population. It is important to collate a more comprehensive set of data, allowing future research to implement PSM with larger sample sizes; to provide more detailed sub-sample analysis; and to permit the use of quantile regression techniques. Furthermore, the inclusion of a question ascertaining complete speaking and writing fluency will provide the basis to compare the relative wage premiums for differing language skills. Consequently, suitable policy suggestions can be made for more targeted government investment in the facilitation of language learning for UK immigrants and ethnic minorities.

Finally, despite reduced subjectivity of the independent variable compared to prior literature, it cannot be ignored that the language variable of interest is not based on an objective test measure, thus is more likely to be subject to measurement error (Nicoletti and Pryor, 2006).

9. Concluding remarks

The prior analysis has shown that there is a significant positive impact of achieving complete reading fluency on the earnings of UK immigrants and ethnic minorities, for whom English is not their first language. OLS analysis demonstrates that the wage premium of achieving complete reading fluency in English is 30.2%. IV estimation addresses the issue of endogeneity and finds a premium of 41.3%. Both econometric methods substantiate the claim of hypothesis 1. The PSM framework addresses the problem of self-selection and confirms the claim of hypothesis 2. The key finding from this approach is that reading deficiencies constrain earnings potential, due to a negative difference in average earnings between non-fluent and fluent individuals by £204 per month. Moreover, the earnings constraint has been proven to exist within industries as well as between, with a mean difference of earnings of £154 and £231 when controlling for high and low skilled industries respectively.

The cross-sectional data taken from the sixth wave of *Understanding Society* via the usage of the Immigrant and Ethnic Minority Boost sample (IEMB) has advantages and limitations. Low sample size and the nature of survey design signify that all prior analysis can only be limited to those residing in areas with a high proportion of immigrants. Moreover, there does not exist an appropriate variable to measure speaking fluency nor an objective measurement for reading proficiency, limiting the scope of the prior analysis. However, significant advantages include the range of nationalities present, which have not previously been analysed in UK literature, and the focus upon immigrants and ethnic minorities facilitating a high-quality matching process.

Significant and large wage premiums for obtaining language proficiency indicates the necessity for the UK government to continue investing in the facilitation of language capital for the current UK immigrant and ethnic minority labour force (Dustmann and Van Soest, 2002). The demand for language courses outstrips supply in high immigrant areas such as London (National Learning and Work Institute, 2017). As such, expanding integration and language programmes for immigrants and ethnic minorities residing in these areas would vastly increase the earnings potential of these individuals, providing positive benefits to the local and national economy via higher spending and taxation payments. For future immigrant workers, given that reading deficiencies constrain earnings potential, an immigration policy that increases the minimum language requirements for those wishing to obtain a Tier 2 general visa should be considered. It is imperative that all four key skills, namely reading, writing, speaking and listening are examined further in future research to provide accurate advice for UK policy makers.

10. Appendix

Figure 7: Dataset by continent

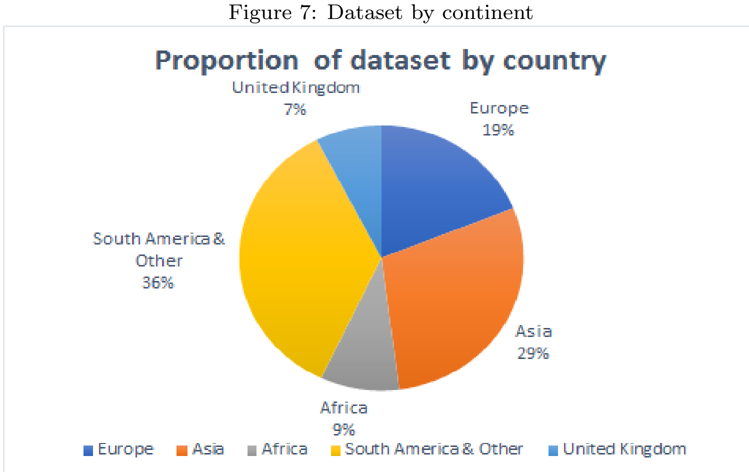


Figure 8: Difference in means test

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	101	6.660224	.0459167	.4614575	6.569126	6.751321
1	646	7.020015	.0270699	.688023	6.96686	7.073171
combined	747	6.971369	.0246254	.6730431	6.923026	7.019712
diff		-.3597916	.070848		-.4988771	-.2207061

diff = mean(0) - mean(1) t = -5.0784

Ho: diff = 0 degrees of freedom = 745

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 1.0000

Figure 9: OLS control variables by group

Personal demographic	Family demographic	Labour market
Age (age)	Married or not (married)	Private firm or not (privfirm)
Age squared (agesq)	Number of children (nochildren)	Highest qualification obtained (highqual)
Gender (f_sex)	Mother's highest qualification (mothhighqual)	Occupation type/industry (occtype)
Years of residence in UK (yrsresuk)	Father's highest qualification (fathhighqual)	Whether highest qualification obtained in the UK or not (f_higheduk)
Disabled or not (f_health)		Whether reside in London or not (London)
Continent of birth (continent)		

Figure 10: Full OLS regressions

VARIABLES	OLS(1) Log of income	OLS (2) Log of income	OLS (3) Log of income	OLS (4) Log of income	OLS Dustmann & Fabri (2003)
f_engread	0.369***	0.506***	0.452***	0.264***	0.276***
	(0.0619)	(0.0612)	(0.0635)	(0.0741)	(0.0692)
age		0.0981***	0.120***	0.0856***	
		(0.0157)	(0.0166)	(0.0167)	
agesq		-0.00104***	-0.00133***	-0.000880***	
		(0.000193)	(0.000205)	(0.000206)	
f_sex		0.249***	0.257***	0.255***	
		(0.0532)	(0.0526)	(0.0502)	
yrsresuk		-0.00606*	-0.00410	-0.00641*	
		(0.00322)	(0.00322)	(0.00355)	
f_health		0.226**	0.199*	0.193*	
		(0.110)	(0.112)	(0.106)	
2.continent		-0.154**	-0.127*	-0.205***	
		(0.0673)	(0.0666)	(0.0631)	
3.continent		-0.230**	-0.174*	-0.247***	
		(0.107)	(0.103)	(0.0948)	
4.continent		-0.177***	-0.148**	-0.174***	
		(0.0597)	(0.0604)	(0.0592)	
married			0.160***	0.137***	
			(0.0576)	(0.0521)	
nochildren			-0.129***	-0.102***	
			(0.0317)	(0.0292)	
2.mothhighqual			0.162**	0.0624	
			(0.0675)	(0.0643)	
3.mothhighqual			0.306***	0.148	
			(0.106)	(0.104)	
2.fathhighqual			-0.0364	-0.0284	
			(0.0636)	(0.0594)	
3.fathhighqual			-0.0577	-0.146*	
			(0.0938)	(0.0841)	
privfirm				0.142**	
				(0.0650)	
2.highqual				-0.0526	
				(0.0718)	
3.highqual				0.104	
				(0.0720)	
4.highqual				0.153*	
				(0.0796)	
2.occtype				0.0284	
				(0.0772)	

Figure 11: OLS output continued

f_higheduk				0.0899	
				(0.0612)	
london				0.180***	
				(0.0476)	
Constant	6.675***	4.339***	4.006***	4.423***	4.538***
	(0.0535)	(0.339)	(0.335)	(0.346)	(0.351)
Observations	554	554	554	554	554
R-squared	0.037	0.184	0.233	0.376	0.254

Note: Controls were added based on proof from prior literature on the variable's impact on earnings. Dustmann and Fabbri controls not shown as different to this paper's specification.

Default categories for dummy variables as follows:

Continent – United Kingdom

f_sex – Female

f_health – not disabled

Married – not married

Mothhighqual, fathhighqual – GCSE level or below

privfirm – not private firm

Occtype – low skilled

Highqual – GCSE level or below

f_higheduk – highest qualification obtained outside UK

London – not residing in London

Figure 12: Tests for instrumental relevance and exogeneity

```
.
. test f_engspk spkengprev

 ( 1)  f_engspk = 0
 ( 2)  spkengprev = 0

       F(  2,   529) =   45.31
            Prob > F =    0.0000

. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  1.72353  (p = 0.1892)
  Basmann chi2(1)        =  1.65089  (p = 0.1988)
```

Figure 13: Language variable additions - logistic regression

Determinants of reading proficiency
Whether spoke English prior to entering uk (spkengprev)
Usual language spoken at home (binary) (langathome)

Figure 14: Logistic regression - mathematical representation

$$\Pr(nonproficient) = F(\alpha + \beta_n x_n + \theta_m x_m + \delta_o x_o + \partial_p x_p)$$

Whereby β, θ, δ, ∂ represent coefficient values for groupings of personal demographic, family demographic, labour market and language variables.

Figure 15: Kernel density distribution - pre and post matching

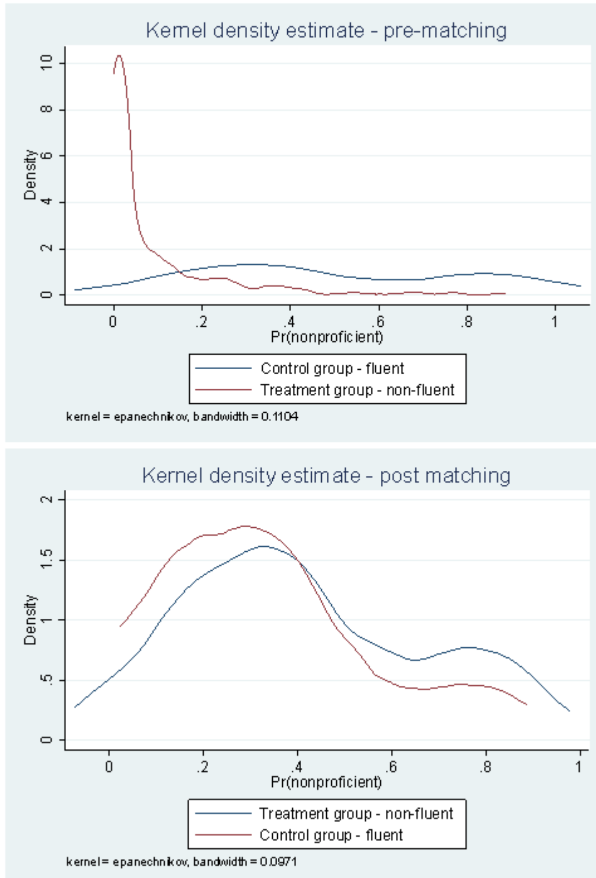


Figure 16: Rosenbaum bounds - Nearest neighbour matching

```
. rbounds diff, gamma(1(0.1)2)

Rosenbaum bounds for diff (N = 67 matched pairs)

Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
----------------------------------------------------------------------
    1        .003718   .003718  -.227869  -.227869  -.413539  -.063365
  1.1        .001268   .009628  -.260939  -.202733  -.442547   -.02903
  1.2        .000423   .020759  -.287964  -.179687  -.471244  -.006122
  1.3        .000139   .038889  -.311711  -.158767  -.501808   .018997
  1.4        .000045   .065265  -.336762  -.136188   -.53103   .043919
  1.5        .000014    .10035  -.359845  -.112166  -.554066   .067744
  1.6        4.5e-06    .14378  -.377932   -.08804  -.573018   .088775
  1.7        1.4e-06   .194471  -.401951  -.072929  -.597498   .106717
  1.8        4.5e-07   .250837   -.41886  -.054746  -.618792   .132047
  1.9        1.4e-07   .311025  -.429767  -.038548  -.638892   .149964
    2        4.3e-08    .37314  -.448364  -.025569  -.655962   .166906

* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)
```

Figure 17: Rosenbaum bounds - Kernel matching

```
. rbounds diff, gamma(1(0.1)2)

Rosenbaum bounds for diff (N = 67 matched pairs)

Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
----------------------------------------------------------------------
    1        .003718   .003718  -.227869  -.227869  -.413539  -.063365
  1.1        .001268   .009628  -.260939  -.202733  -.442547   -.02903
  1.2        .000423   .020759  -.287964  -.179687  -.471244  -.006122
  1.3        .000139   .038889  -.311711  -.158767  -.501808   .018997
  1.4        .000045   .065265  -.336762  -.136188   -.53103   .043919
  1.5        .000014    .10035  -.359845  -.112166  -.554066   .067744
  1.6        4.5e-06    .14378  -.377932   -.08804  -.573018   .088775
  1.7        1.4e-06   .194471  -.401951  -.072929  -.597498   .106717
  1.8        4.5e-07   .250837   -.41886  -.054746  -.618792   .132047
  1.9        1.4e-07   .311025  -.429767  -.038548  -.638892   .149964
    2        4.3e-08    .37314  -.448364  -.025569  -.655962   .166906

* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)
```

Figure 18: Rosenbaum bounds - Radius matching

```
. rbounds diff, gamma(1(0.1)2)

Rosenbaum bounds for diff (N = 67 matched pairs)

Gamma           sig+      sig-    t-hat+    t-hat-       CI+       CI-
----------------------------------------------------------------------
    1        .003718   .003718  -.227869  -.227869  -.413539  -.063365
  1.1        .001268   .009628  -.260939  -.202733  -.442547   -.02903
  1.2        .000423   .020759  -.287964  -.179687  -.471244  -.006122
  1.3        .000139   .038889  -.311711  -.158767  -.501808   .018997
  1.4        .000045   .065265  -.336762  -.136188   -.53103   .043919
  1.5        .000014    .10035  -.359845  -.112166  -.554066   .067744
  1.6        4.5e-06    .14378  -.377932   -.08804  -.573018   .088775
  1.7        1.4e-06   .194471  -.401951  -.072929  -.597498   .106717
  1.8        4.5e-07   .250837   -.41886  -.054746  -.618792   .132047
  1.9        1.4e-07   .311025  -.429767  -.038548  -.638892   .149964
    2        4.3e-08    .37314  -.448364  -.025569  -.655962   .166906

* gamma  - log odds of differential assignment due to unobserved factors
  sig+   - upper bound significance level
  sig-   - lower bound significance level
  t-hat+ - upper bound Hodges-Lehmann point estimate
  t-hat- - lower bound Hodges-Lehmann point estimate
  CI+    - upper bound confidence interval (a=  .95)
  CI-    - lower bound confidence interval (a=  .95)
```

11. References

Becker, S. and Caliendo, M. (2007). mhbounds Sensitivity Analysis for Average Treatment Effects. Discussion Paper No. 2542. [online] Berlin: Marco Caliendo. Available at: http://ftp.iza.org/dp2542.pdf [Accessed 27 Mar. 2018].

Borjas, G. (1994). The Economics of Immigration. Journal of Economic Literature, [online] 32, p.1684.

Budria, S. and Swedberg, P. (2015). The impact of language proficiency on immigrant's earnings. Journal of Applied Economics, [online] 23(67), pp.63 - 91.

Bleakley, H. and Chin, A. (2004). Language Skills and Earnings: Evidence from Childhood Immigrants*. Review of Economics and Statistics, 86(2), pp.481-496.

British Council (2013). The English Effect. [online] British Council, p.17. Available at: https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf [Accessed 30 Nov. 2017].

Caliendo, M. and Kopeinig, S. (2008). SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. Journal of Economic Surveys, 22(1), pp.31-72.

Chiswick, B. (1978). The Effect of Americanization on the Earnings of Foreign-born Men. Journal of Political Economy, 86(5), pp.897-921.

Chiswick, B and Miller, P. (1990). Language in the Labor Market: The Immigrant Experience in Canada and the United States. Canada/United States.

Chiswick, B. (1991). Speaking, Reading, and Earnings among Low-Skilled Immigrants. Journal of Labor Economics, 9(2), pp.149-170.

Chiswick, B. and Miller, P. (1995). The Endogeneity between Language and Earnings: International Analyses. Journal of Labor Economics, 13(2), pp.246-288.

Cochran, W. and Rubin, D. (1973). Controlling Bias in Observational Studies: A Review. The Indian Journal of Statistics,p.p.420.

Dehejia, R. and Wahba, S. (2002). Propensity Score-Matching Methods for Nonexperimental Causal Studies. Review of Economics and Statistics, 84(1), pp.151-161.

Di Paolo, A. and Raymond, J. (2012). LANGUAGE KNOWLEDGE AND EARNINGS IN CATALONIA. Journal of Applied Economics, 15(1), pp.89-118.

Dustmann, C. (1994). Speaking fluency, writing fluency and earnings of migrants. Journal of Population Economics, 7(2).

Dustmann, C. and Van Soest, A. (2002). Language and the Earnings of Immigrants. ILR Review, 55(3), pp.473-492.

Dustmann, C. and Fabbri, F. (2003). Language proficiency and labour market performance of immigrants in the UK*. The Economic Journal, 113(489), pp.695-717.

Ft.com. (2017). UK proposal threatens tough post-Brexit rules for EU migrants. [online] Available at: https://www.ft.com/content/d88c72d8-9265-11e7-bdfa-eda243196c2c [Accessed 30 Nov. 2017].

Ginsburgh, V. and Prieto-Rodriguez, J. (2013). Is there a Gender Bias in the Use of Foreign Languages in Europe?. Kyklos, 66(4), pp.552-566.

Gov.uk. (2017). Preparing for adult funding devolution - GOV.UK. [online] Available at: https://www.gov.uk/government/news/preparing-for-adult-funding-devolution [Accessed 28 Mar. 2018].

Grenier, G. (1984). The Effects of Language Characteristics on the Wages of Hispanic-American Males. The Journal of Human Resources, 19(1), p.35.

Guven, C. and Islam, A. (2015). Age at Migration, Language Proficiency, and Socioeconomic Outcomes: Evidence From Australia. Demography, 52(2), pp.513-542.

Hamilton, T. (2014). Selection, Language Heritage, and the Earnings Trajectories of Black Immigrants in the United States. Demography, 51(3), pp.975-1002.

Heckman, J., Lalonde, R. and Smith, J. (1999). THE ECONOMICS AND ECONOMETRICS OF ACTIVE LABOR MARKET PROGRAMS. Handbook of Labour Economics, 3, part A.

Hwang, S., Xi, J. and Cao, Y. (2010). The conditional relationship between English language proficiency and earnings among US immigrants. Ethnic and Racial Studies, 33(9), pp.1620-1647.

Isphording, I. (2013). Returns to Foreign Language Skills of Immigrants in Spain. LABOUR, 27(4), pp.443-461.

Lechner, M. (2008). A Note on the Common Support Problem in Applied Evaluation Studies. Annales d'conomie et de Statistique, (91/92), p.217.

McManus, W., Gould, W. and Welch, F. (1983). Earnings of Hispanic Men: The Role of English Language Proficiency. Journal of Labor Economics, 1(2), pp.101-130.

National Learning and Work Institute (2017). Mapping ESOL Provision in Greater London. London: National Learning and Work Institute.

Nicoletti, G. and Pryor, F. (2006). Subjective and objective measures of governmental regulations in OECD nations. Journal of Economic Behavior & Organization, 59(3), pp.433-449.

Office for National Statistics (2014). Detailed Analysis - English Language Proficiency in the Labour Market. Office for National Statistics, p.7.

Office for National Statistics (2017). Population of the UK by country of birth and nationality: 2016. [online] Office for National Statistics, p.5.

Office for National Statistics (2018). Migration Statistics Quarterly Report: February 2018. [online] Office for National Statistics.

Reimers, C. (1983). Labor Market Discrimination Against Hispanic and Black Men. The Review of Economics and Statistics, 65(4), p.570.

Rivera-Batiz, F. (1990). English language proficiency and the economic progress of immigrants. Economics Letters, 34(3), pp.295-300.

Rosenbaum, P. (2005). Observational Study. Encyclopedia of Statistics in Behavioral Science, [online] Volume 3, pp.pp. 14511462. Available at: http://www-stat.wharton.upenn.edu/ rosenbap/BehStatObs.pdf [Accessed 28 Mar. 2018].

A. Smith, J. and E. Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. Journal of Econometrics, 125(1-2), pp.305-353.

Tainer, E. (1988). English Language Proficiency and the Determination of Earnings among Foreign-Born Men. The Journal of Human Resources, 23(1), p.108.