ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Budría, Santiago; Martínez de Ibarreta, Carlos

Working Paper

Educational and Skills Mismatches among Immigrants: The Impact of Host Language Proficiency

IZA Discussion Papers, No. 13030

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Budría, Santiago; Martínez de Ibarreta, Carlos (2020) : Educational and Skills Mismatches among Immigrants: The Impact of Host Language Proficiency, IZA Discussion Papers, No. 13030, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/216342

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13030

Educational and Skills Mismatches among Immigrants: The Impact of Host Language Proficiency

Santiago Budría
Carlos Martínez de Ibarreta

MARCH 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13030

Educational and Skills Mismatches among Immigrants: The Impact of Host Language Proficiency

Santiago Budría
Universidad Antonio de Nebrija, CEEAplA and IZA

Carlos Martínez de Ibarreta
Universidad Pontificia Comillas

MARCH 2020

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Educational and Skills Mismatches among Immigrants: The Impact of Host Language Proficiency*

This paper asks to what extent host language proficiency can insure immigrants against the risk of ending up in mismatched jobs. Using the 2003-2016 waves of the Household, Income and Labour Dynamics in Australia (HILDA), the paper discriminates between three forms of mismatch, overqualification, under-qualification and over-skilling. Host language proficiency is instrumented using Bleakley and Chin (Rev Econ Stat 86:481–496, 2004) strategy, which exploits the fact that younger children learn languages more easily than older ones. To differentiate between local average treatment effects (LATE) and average treatment effects (ATE), the paper considers two alternative models, 2SLS instrumental variables and biprobit. We find that treatment effects are heterogeneous. English language proficiency among immigrants in Australia reduces the probability of ending up in over-qualified jobs, by between 17.2 (LATE) and 36.7 (ATE) percentage points. The ATE of overs-skilling is also significant and about -8.9 percentage points. In contrast, language skills tend to raise the probability of being under-qualified at the job, by about 8.6 percentage points according to the ATE. Local effects of over-skilling and underqualification fail to be statistically significant, suggesting that host language proficiency may be innocuous for some workers. Overall, the results indicate that host language proficiency is a country-specific, valuable form of human capital.

JEL Classification: F22, J24, J61

Keywords: over-qualification, under-qualification, over-skilling, host language proficiency, instrumental variables

Corresponding author:
Santiago Budría
Department of Business Administration
Universidad Nebrija
C/ de Sta. Cruz de Marcenado, 27
28015 Madrid
Spain
E-mail: sbudria@nebrija.es

* The author gratefully acknowledges the financial support provided by the R&D Program in Social Sciences and Humanities by the Autonomous Community of Madrid, OPINBI project (Ref: H2019/HUM-5793).

1. Introduction and background

Understanding the factors that determine the economic performance of immigrants is crucial to support their process of assimilation. The role of fluency in the host country's dominant language has been a prominent theme in the literature. Host language may enhance productivity on the job by making the worker more efficient in performing particular tasks and/or by reducing the cost of communication within the firm. Besides, proficient workers are in a better position to obtain information about job opportunities, and to transmit valuable information about their skills and background to employers. Consistent with this, the literature has shown that host language proficiency is a relevant determinant of labour market performance. Most of the interest has gravitated around a particular aspect of the labour market, earnings (Chiswick and Miller, 2010, Zhen, 2013, among many others), while evidence on other labour market outcomes, including employment and occupational attainment (Aldashev et al., 2009, Yao and Van Ours, 2015) is relatively more scarce.

This paper puts the focus on an important yet unexplored labour market condition, the quality of the match between the education and skills possessed by the individual and the requirements of the job. We use the 2003-2016 waves of the HILDA dataset, a micro panel survey representative of the Australian population, to examine the role of host language proficiency in easing immigrant's access to matched jobs. At the aggregate level, misallocations of workers to jobs may imply significant productivity losses and negatively affect long-run growth (Hsieh et al., 2019). Relative to comparable matched individuals, workers with excess qualifications and skills are subject to lower earnings, job satisfaction and employment opportunities (Verhaest and Omey, 2009, Mavromaras et al. 2012, McGuinness et al., 2018, to cite a few). These effects might pose an additional disadvantage to immigrants, an already disfavoured group. In contrast, low-educated workers holding higher level jobs than would be expected based on their formal schooling (the under-qualified) tend to benefit from increased personal development, job engagement and, in some cases, earnings and subjective well-being, relative to matched individuals with similar levels of schooling (van der Velden and Verhaest, 2017).

To accommodate this evidence, we discriminate between two forms of educational mismatch, over-qualification and under-qualification. Additionally, we consider skills mismatches, namely over-skilling. Educational and skill mismatches refer to quite different phenomena. Workers with excess qualifications may still lack skills that are necessary on the job, while workers lacking qualifications may possess skills that are needed on the job. Moroever, over- and under-qualification are closely related to education achievement, while skills mismatches are more prone to capture work-related human capital. From an individual

point of view, the determinants of education and skills mismatches are found to differ, and the correlation between these two indicators is weak (Flisi et al., 2017).

The paper is concerned with causal effects. The potential endogeneity of host language proficiency is controlled for by using information on the language spoken during childhood and age at arrival in Australia as a source of external variation. Being born in a non-English speaking country implies a worse command of English at adulthood, while early age at arrival in the host country can make up for this initial disadvantage. Because age at arrival only affects language skills of immigrants whose mother tongue is not English, we follow Bleakley and Chin (2004) and take the interaction between these two variables as identifying instrument.

We present results from two econometric models, 2SLS instrumental variables and a Bivariate Probit (BP). It is known that 2SLS captures the local average treatment effect (LATE), implying in our case the effect on immigrants who are induced into the treatment by the instrumental variable (the complier subpopulation). Under 2SLS, the corresponding treatment effects for non-compliers are, by definition, not identified. The focus of 2SLS analysis on compliers raises questions about how different this group is from the non-compliers, and by how much the LATE differs from the ATE, the average treatment effect (Heckman and Urzua, 2010). The latter is often the main quantity of interest for applied researchers but different from the LATE if treatment effects are heterogeneous. Therefore, we extend our analysis to include maximum-likelihood estimates of a BP, which assumes that the outcome and treatment are each determined by latent linear index models with jointly normal error terms. BP results are then used to infer the ATE and the average effect of treatment on the treated (ATET). The simultaneous consideration of two different models is reinforced by the fact that 2SLS and BP estimators are found to differ in their sensitiveness to model misspecification, departure from the assumption of normality and the width of confidence intervals (Chiburis et al., 2012).

We find evidence that host language proficiency reduces the probability of being over-qualified, by between 17.2 and 36.7 percentage points (pp) and, to a lower extent, the probability of over-skilling. In contrast, language skills increase the likelihood of ending up in under-qualified jobs, implying that for some immigrants in the Australian labour market language skills make up for insufficient formal schooling. Moreover, we document heterogeneous treatment effects between LATE and ATE estimates. Finally, we show that the results pass well a battery of sensitivity checks, such as accounting for methodological changes in the definition of education and skills mismatches, the use of very detailed industry-occupation cells when appraising schooling requirements, assessing the validity of the instrument and controlling for attrition and transitions to and out of the labour market.

The paper is organized as follows. Section 2 provides a brief background of the literature. Section 3 describes the dataset, the estimating sample and the English language proficiency question. Section 4 introduces the 2SLS and BP approaches and the corresponding

LATE, ATE and ATET estimators. Section 5 presents the results for the impact of English language proficiency on the various forms of mismatch. Section 6 contains a battery of sensitivity checks and presents additional estimates. Section 7 contains the concluding remarks. The paper includes an Appendix with supplementary tables and results.

2. Background and previous literature

Educational and skills mismatches have consequences on a variety of labour market indicators, including earnings, job satisfaction and employment transitions. These associations have been documented extensively under a myriad of datasets, countries, econometric approaches, definitions and measurement methods (McGuinness et al., 2018, for a recent survey).

Despite the accumulated evidence, there is still not much consensus about the main causes of mismatch. Several theories in the literature support the view of the phenomenon being a short-term problem. Matching theory (Jovanovic, 1979) suggests that over-qualification represents a poor match for workers because they are qualified to perform higher level jobs. Over time, however, workers are expected to improve their job match. In the same line, the career mobility hypothesis supports that over-qualification is part of a career path or insertion process in the labour market (Büchel and Mertens, 2004).

Other theories consider job mismatches as a more serious and long-lasting problem. This occurs, for instance, when the labour market is characterized by imperfect information – Spence's (1973) job-screening model – or when the presence of labour market rigidities induces workers to occupy jobs for which they are over-qualified. Assignment models (Sattinger, 1993) also stress the importance of job distribution. Under this setting, the job allocation process and utility maximization guides workers to choose certain jobs over others. Therefore, changes in the distribution of earnings and, by default, over- and under-qualification, will be related to both the distribution of jobs and the characteristics of the workforce.

Complementary explanations focus on the presence of unobserved heterogeneity. Under-qualified workers may be in some way more able and possess abilities and skills that allow them to access jobs that require formally more schooling than they possess. Results pointing to a wage premium among under-qualified workers relative to workers with the same education in matched jobs are consistent with this view (Verhaest and Omey, 2012). Similarly, workers with excess qualifications may lack some of the abilities and skills required to do a job that is not aligned with their level of education, although the evidence supporting this hypothesis is limited. For instance, the wage penalty associated with over-qualification is independent of the level of skill utilization within firms and the skills and abilities possessed by the individual (McGuiness and Bennet, 2007).

Finally, studies investigating the correlates of educational mismatches typically find that over-qualification is more prevalent among the high-educated, part-time workers and individuals with previous experiences of over-qualification (Belfield, 2010, Verhaest et al., 2015), whereas skills mismatches are poorly related to schooling levels and show some persistence over time (Mavromaras and McGuinness, 2012).

2.1. Measurement

Over-qualification describes the extent to which an individual possesses a level of education in excess of that which is required for her job. Three ways of measuring this phenomenon are commonly accepted in the literature: the respondents' subjective assessments, realized matches within occupational classifications and job analysts' ratings. The choice of a particular method in applied work depends mostly on data availability. The subjective approach is based on the worker's self-assessment regarding the quality of the match between her education and the educational requirements of the job, while in the other two methods educational requirements are appraised objectively. Realized matches is a statistical procedure according to which the job-level requirements are given by a central measure (mean, modal or median) of years of schooling within the occupation. A worker is regarded to be over-qualified (under-qualified) if she has more (less) education than is required for the job. Under the job analysts method, required schooling is established by experts.

All methods have specific advantages and limitations. Workers' self-assessments may have a tendency to overstate the requirements of the jobs and to upgrade the worker's status of her position. Moreover, perceptions of the requirements of the job may be biased, and this bias can be driven by socio-economic factors (Leuven and Oosterbeek, 2011). An additional concern is that subjective indicators are typically a binary variable that prevents researchers from measuring the intensity of over- and under-qualification. A potential advantage of self-assessment is that workers have privileged information about their background, job requirements and hiring standards of the firm. In contrast, realized matches may reflect average credentials of all workers within a given occupation and, therefore, may be more closely related to education levels required "to get" as opposed "to do" a given job in contemporary terms. In exchange, realized matches capture time trends in the educational distribution within occupations and allow researchers to measure the intensity of the mismatch. Sensitivity analyses show that the extent, effects and determinants of over- and under-qualification may differ across measures, although differences are moderate (European Commission, 2015).

As for the measurement of skills mismatches, most studies rely on self-reported data. This is due to the difficulty of defining and surveying all skills that individuals possess and the job requires. The Reflex dataset has been probably the most utilized source to investigate the determinants and consequences of skills mismatches (Baert at al., 2013). In this dataset, the

central question is “to what extent are your skills utilized in this work”, the possible answers ranging from of 1 (not at all) to 5 (to a very high extent). Low answers are then interpreted as an indicator of over-skilling. A separate question with an identical range of possible answers is intended to measure under-skilling, “to what extent does this work require more knowledge and skills than you can actually offer”. In the HILDA dataset, respondents are asked to state their level of agreement with the statement: “I use many of my abilities in my current job”. Responses to this statement take the values of 1 (strongly agree) to 7 (strongly disagree) and are the basis of studies on over-skilling in Australia (Mavromaras et al. 2007, Zhu and Chen, 2016).

2.2. The role of host language proficiency among immigrants

The primary divergence in the labour market between immigrant and native workers stems from the access to jobs and the quality of them (Chiswick and Miller, 2009). Immigrants are at a disadvantage when they enter the labour market due to less-than-perfect transferability of skills and reduced labour market information. They are also more vulnerable to cyclical, seasonal and transitory changes in economic activity and, therefore, experience more involuntary job turnover and unemployment (Chiswick and Hurst, 2000). This evidence is suggestive of a secondary labour market for immigrants in which most of them are limited to work in certain sectors of the economy and end up occupying “immigrant jobs”. This idea is consistent with the traditional view in labour economics that the labour market reward to worker’s characteristics is frequently occupation-dependent.

In this context, host language proficiency may act as a buffering component, reducing barriers and improving access to jobs. The empirical evidence is consistent with this view. For instance, there is extensive evidence showing that host language skills are positively correlated with earnings (Dustmann and Van Soest, 2001, Chiswick and Miller, 2010, Zhen, 2013). Although the estimates differ sensitively across countries, languages and datasets, the accumulated evidence suggests that ceteris paribus proficient immigrants reap an earnings premium of between 5% and 20%. However, language proficiency can be viewed as having both direct and indirect impacts on earnings, with the indirect impacts operating via occupational attainment. For instance, Chiswick and Miller (2009) estimate that among the foreign born in the US, between 11.3% and 48.7% of the estimated lower earnings due to limited English-language proficiency occurs because they are placed in lower earnings occupations (inter-occupational differences). Shields and Price (2001) also find that English language speaking fluency is clearly an important determinant of employment for both male and female immigrants in the UK. They report that speaking English fluently increases the average predicted probability of employment by 20-25 pp, relative to an immigrant with poor English-speaking ability. Dustmann and Fabbri (2003) point out that language may help to acquire information about optimal job search strategies. Moreover, many jobs, especially those in the

service sector, require communication skills. Consistent with this idea, they find that fluency in English increases employment probabilities by about 22 pp. Budría et al. (2019) show that in Spain host language proficiency raises the probability of having a job by between 15 and 22 pp. Yao and van Ours (2015) find that in the Netherlands male immigrants with poor Dutch skills have the same employment probability and hours of work as male immigrants without language problems. This result contrasts with the evidence based on world-wide spoken languages, such as English and Spanish, and suggests that language effects are related with the scope of the language.

Finally, Green et al. (2007) use Australian data to examine the quality of the match between the education possessed by the individual and the requirements of the job. Although their paper does not focus on the role of host language skills, they find that the extent of over-qualification differs largely between immigrants depending on whether they have an English-speaking background. Relative to the native born, immigrants from non-English speaking regions have higher educational qualifications but are 5 times more likely to be over-qualified, and 2 times more likely relative to immigrants from an English-speaking background.

3. Data and definition of variables

The data used in this paper are taken from the 2003-2016 waves of the Household, Income, and Labour Dynamics in Australia (HILDA) Survey. With a yearly structure, the HILDA dataset is Australia's first nationally representative household panel survey. Each wave covers approximately between 7,000 and 9,500 households drawn from 13 areas of Australia and includes about 20,000 individuals. We restrict the sample to full-time wage earners aged between 21 and 60 who are not employed in the agricultural sector. Workers in part-time jobs may have chosen to do so for reasons of family or other personal commitments or preferences (e.g., flexibility of hours of work, shorter distances to work) and, thus, may be more likely to accept mismatched jobs. Therefore, part-time workers are excluded from the analysis to avoid selection bias. The case of women is disregarded on account of the extra complication of endogenous labour market participation. Immigrants account for 26.8% of the resulting sample, a figure that is sensitively below its census counterpart (33.3%). In the regression stage of the paper we retain only immigrants. These restrictions leave us with a final sample of 11,224 observations, from 1,906 individuals across 14 years.

3.1 Educational and skills mismatches

We use information from realized matches to establish job-level school requirements by occupation. Key studies in the literature adopting the realized matches approach include Cohn

and Kahn (1995), Bauer (2002) and Summerfield and Theodossiou (2017). We assume that required schooling is given by the one standard-deviation range around the mean level of schooling within the occupation. People are the defined to be over- or under-qualified depending on whether they are above or below this range. The one standard-deviation range captures the dispersion of schooling levels within an occupation and adapts the range of required schooling accordingly.

The distribution of schooling within occupations is expected to be dynamic due to changes in technology, educational quality and supply and demand forces. To account for time trends in the various occupational assignments, we allow required education to vary with survey year. Still, realized matches are subject to some drawbacks. For instance, the cut-off at one standard deviation from the mean is arbitrary, and the mean can be sensitive to outliers. For this reason, we also consider two variations of this method. These consists in replacing the one standard-deviation range around the mean by either the modal or median schooling level. Robustness checks are provided in Section 6. As a second limitation, the realized matches approach ignores variation in required schooling across jobs within an occupation. For this reason, in Section 6 we report additional results with very detailed occupation-industry cells.

In the HILDA survey education is coded in seven categories.[1] We transform these categories into years schooling following the criterion described in Lillard et al. (2016). Occupations are disaggregated at the 2-digit ANZSCO code level. Groups with fewer than ten observations within each year are excluded because the required level of education generated from such small samples is unlikely to be representative.[2] For each occupation and year, we compute the mean and the standard deviation of years of schooling. To gain representativity, this information is extracted from the full HILDA sample (41,881 obs.), not just the subsample of immigrants. Finally, we classify immigrant workers according to the resulting thresholds. In complementary calculations, we combine the ANZSCO occupations with 2-digit ANZSIC industry codes, generating a 4-digit classification of detailed industry-occupation cells. In Figure 1 we report the distribution of occupation-specific average years of schooling. The average across waves and occupations is 12.7 years, with a minimum of 11.0 years ("Plant and machine operators", ANZSCO #80) and a maximum of 16.2 years ("Legislators", ANZSCO #11).

[1] The education levels are: i) postgrad-masters or doctorate, ii) graduate diploma or graduate certificate, iii) bachelor or honours, iv) advanced diploma or diploma, v) certificate III or IV, vi) year 12 and vii) year 11 and below.

[2] This exclusion restriction affects a yearly average of 0.69% of the workers in the sample.

As for skills mismatches, we use the subjective question described and contextualized in the literature review, "I use many of my skills and abilities in my current job", with possible answers ranging from 1 (strongly disagree) to 7 (strongly agree). Notwithstanding the possible shortcomings of all self-reported information, this question allows us to detect the presence of skills that can be relevant in the employment context and that, perhaps, are not represented by formal educational achievements. We consider respondents selecting 1–4 as over-skilled and use those selecting 5–7 as the reference category (matched). [3]

3.3 English language proficiency

The English language proficiency question on the HILDA is:

- *"How well you speak English?"*

Available answers range from 1 (very well) to 4 (not at all). The responses were used to define *EP,* a dummy variable that takes value one if the immigrant is proficient in English (1-very well), zero otherwise.[4] According to this criterion, nearly 75.9% of the sample reports being proficient in English.

Due to the high costs of test-based assessments of language ability in large scale surveys, the literature to date has relied on self-reported competence. One criticism against the use of subjective rather than objective measures of an individual's language fluency is that it can lead to under or overestimation of the actual language proficiency, since an immigrant is likely to judge his ability relative to a fellow immigrant, and not in comparison with a native-born citizen. Moreover, respondents may have different perceptions under identical circumstances of how well they speak a foreign language. This notwithstanding, subjective questions are typically found to be highly correlated with scores from tests designed to accurately measure language ability as well as functional measures of language skills' (Akbulut-Yuksel et al., 2011).

4. Models and estimators

Let English proficiency, $EP \in [0, 1]$, be a potentially endogenous treatment, and mismatch status $M \in [0, 1]$ be the outcome of interest. Let M_1 be an individual's potential outcome had she received the treatment ($EP = 1$) and let M_0 be the individual's potential outcome had she not

[3] An alternative classification (1-5 against 6-7) leaded to very similar results.

[4] The paper follows a stringent criterion by considering only individuals who claim to be able to speak English 'very well'. Results under the alternative classification 1-2 against 3-4 displayed similar effects and are available upon request.

received the treatment ($EP = 0$). Let EP_1 be an individual's English language knowledge after a one-unit increase in instrument Z ($\Delta_z = 1$) and let EP_0 be an individual's English language knowledge with $\Delta_z = 0$. The ATE over the entire population is given by:

$$\text{ATE} = \text{E}[M_1] - \text{E}[M_0] \tag{1}$$

and represents the expected effect of English proficiency on an immigrant randomly drawn from the population. The ATET is the average treatment effect only among those individuals who actually received the treatment ($EP = 1$):

$$\text{ATET} = \text{E}\,[M_1 \,|EP = 1] - \text{E}\,[M_0 \,|EP = 1] \tag{2}$$

Finally, the LATE is:

$$\text{LATE} = \frac{E[M\ |\Delta_Z=1\ ]-E[M\ |\Delta_Z=0\ ]}{E[T\ |\Delta_Z=1\ ]-E[T\ |\Delta_Z=0\ ]} \tag{3}$$

and focuses on immigrants who are induced into the treatment by the instrumental variable (the complier subpopulation). Therefore, it is uninformative about effects on individuals whose treatment status is not affected by the instrument.

4.1 BP and 2SLS

In the BP model mismatch status is modelled as a function of a latent variable M^* that is not measured, is continuous, has a threshold point that determines the observed value of M, and is a function of observable characteristics, while the treatment is assumed to be endogenous:

$$\begin{aligned} M_i &= I(X_i\alpha + \theta EP_i + \varepsilon_{i1} > 0) \\ EP_i &= I(X_i\delta + \gamma Z_i + \varepsilon_{i2} > 0) \end{aligned} \tag{4}$$

The indicator function $I(\cdot)$ equals one if its argument holds and zero otherwise, and the error terms are jointly distributed as standard bivariate normal with correlation ρ, (ε_{i1} , $\varepsilon_{i2}) \sim B[(0,0),(1,1),\rho]$, $\rho \in [-1,1]$. Parameter ρ captures the correlation between unobservables that affect mismatch status and the likelihood of being English proficient. Language knowledge may depend on unobservable individual characteristics that are potentially related to unmeasurable employment determinants. That would be the case if, for example, more productive and capable individuals were more likely to be proficient in English. Therefore, Z is an instrumental variable that must be *valid* and *relevant*, i.e., it only affects outcome through

treatment, conditional on covariates, and is correlated with treatment *EP*. This model can be estimated using maximum likelihood, and the estimated parameters can be plugged into the formulas for ATE and ATET. Let ϕ be the standard normal distribution. Then, following Chiburis et al. (2012), we can write:

$$\text{ATE} = \phi\,(X_i\alpha + \theta) - \phi\,(X_i\alpha) \tag{5}$$

$$\text{ATET} = \Pr(\Delta_Z = 0)\ \frac{B[X_i\delta+\gamma Z_i, X_i\alpha+\theta,\ \rho]-B[X_i\delta+\gamma Z_i, X_i\alpha,\ \rho]}{\phi\,(X_i\delta+\gamma Z_i)} + \Pr(\Delta_Z = 1)\frac{B[X_i\delta+\gamma(Z_i+1), X_i\alpha+\theta,\ \rho]-B[X_i\delta+\gamma(Z_i+1), X_i\alpha,\ \rho\,]}{\phi\,(X_i\delta+\gamma(Z_i+1))} \tag{6}$$

In the regressions stage, we calculate estimators $\widehat{ATE}$ and $\widehat{ATET}$ and their standard deviations by averaging across observations and performing 500 bootstrap replications.

Using 2SLS instead of BP involves ignoring the binary dependent nature of the outcome and treatment variables and replacing Eq. (4) by its linear functional forms. Under regular assumptions of instrument validity and relevance, it can be shown that $\widehat{LATE} = \hat{\theta}$ (Angrist and Pischke, 2009).

4.2 Covariates

Vector *X* includes socio-economic factors that are standard when accounting for labour market outcomes among immigrants, including years of schooling, age, age at arrival, marital status (married, divorced/widowed, reference: single), dummy variables for parenthood and previous unemployment experience, working hours, tenure at the firm, sector of activity (private, non-profit, reference: public sector), firm size and region of residence (there are 8 Australian regions). Since origin is related to both job opportunities and English skills, we also control for the worker's source geographical region (Oceania, Southern Europe, Eastern Europe, Maghreb and Middle East, South-East Asia, East Asia, South and Central Asia, North America, Latin-America and Sub-Saharan Africa, reference: Northern Europe).

Mismatch rates are also related to structural imbalances between the overall demand and supply of skilled workers and differences in the business cycle. Specifically, over-qualification rates have been found to be negatively correlated with employment rates, economic growth and institutional factors including labour market flexibility, family conciliation support and the orientation of national study programs towards vocational training and the development of specific skills (Verhaest and van der Velden, 2013, Summerfield and Theodossiou, 2017). Therefore, we also include variables to control for demand and business cycle effects at the regional level. Specifically, the proportion of female workers over total employment and the regional participation rate are intended to capture the extent of competition

for jobs in the labour market and relative demand effects. The labour force share of part-time workers controls for the fact that regions with higher employment shares of temporary and/or part-time workers have generally an increased capacity to respond to labour market disequilibria. The stance of the business cycle also influences the extent to which individuals manage to get a good match. To account for general macroeconomic conditions, we include the regional unemployment rate, per capita GDP and GDP growth. We also include the percentage of immigrant people over the total population.[5] Stigma and segregation effects among immigrants may be present at the labour market and these may depend on the relative size of the immigrant population. Finally, workers in regional labour markets with a higher (lower) proportion of educated individuals are, ceteris paribus, more likely be over-qualified (under-qualified) at their jobs. Hence, we include the regional share of workers with university education in the equation.

4.3 Selected Instrument

Age at arrival is negatively correlated with language knowledge, since younger children learn languages more easily than adolescents and adults. Cognitive scientists refer to this as the *critical period hypothesis* according to which there is a critical age range in which individuals learn languages more easily (Chiswick et al., 2008). Figure 2 shows the relationship between age at arrival and English-language skills among immigrants with English-speaking background (ESB) and immigrants with non-English-speaking background (NESB). We classify immigrants as ESB or NESB depending on whether they were born in a country where English is an official language. ESB immigrants are essentially all fluent in English, regardless of their age at arrival. This is not surprising, insofar as their first exposure to English does not depend on when they migrated to Australia. Consistent with the *critical period hypothesis*, NESB immigrants who received their first exposure to English at an early age attain English language skills comparable to those of ESB immigrants. By contrast, immigrants whose first exposure to English was after 9 exhibit lower skills, and the disadvantage increases almost linearly with age at arrival.

-Insert Figure 2 here-

However, age at arrival itself cannot be an instrument, since early arrival fosters better knowledge of the host society, cultural convergence and, therefore, may lead to employment in a matched job. Still, by including age of arrival as a control variable in the mismatch status equation we can partial out the non-language effects of early arrival. This occurs because upon

[5] This percentage is obtained from the census of born overseas, and is available for the years 2001, 2006, 2011, 2016. The figures included in the regressions are a rolling average across years.

arrival in Australia, ESB immigrants experience everything that NESB immigrants encounter, except for learning a new language. Therefore, we adopt the following parameterization for the instrument:

$$Z_i = \max(0, \text{age at arrival}_i - 9) \times \text{I(NESB immigrant}_i) \quad (7)$$

where I() is the indicator function. This formulation closely follows Bleakley and Chin's (2004) identification strategy and captures much of the co-movement between age at arrival and English-language skills displayed in Figure 2. Robustness checks with alternative age cut-off points are provided in Section 6.

The bias of a 2SLS estimator is given by:

$$Bias_{2SLS} = \frac{\sigma_{EP,\varepsilon_1}}{\sigma^2_{\varepsilon_2}} \frac{1}{F+1} \quad (8)$$

where $\sigma_{EP,\varepsilon_1}$ is the correlation between the endogenous variable and the error term of the second stage equation, $\sigma^2_{\varepsilon_2}$ is the residual variance in the first stage equation and F is the statistic for the significance of excluded instruments (see, Angrist and Pischke, 2009, p: 208). Therefore, having relevant instruments is very important to attenuate any potential bias. Stock et al.'s (2002) influential work suggests that F-statistics above about 10 keep the extent of bias to the safe zone. As we shall see, the selected instrument passes well this threshold. [6]

Table A1 in the appendix shows the decomposition of the sample by country of birth and presents the classification of countries by English-speaking status.

5. Results

5.1 Preliminary evidence

In Table 1 we present the incidence of educational and skills mismatches by immigrant status. The immigrant over-qualification rate, above 20%, appears very high when compared to the native population, 11.1%. This gap is in line with the figures reported in Green et al. (2007), despite the fact that they use a different dataset (the Longitudinal Survey of Immigrants to Australia) and job analysis data to determine the occupational requirements of jobs. Therefore, there is evidence to suggest that over-qualification poses an additional disadvantage to immigrants, an already disfavoured group, especially if they are not English proficient.

[6] In addition, the model is just-identified, because the interaction term is the only instrument. Just-identified 2SLS is approximately unbiased, even with weak instruments.

Similarly, the incidence of over-skilling is larger among the non-proficient than among native and proficient immigrants. By contrast, the incidence of under-qualification is slightly higher among native workers. This pattern may be indicative of native individuals finding it easier to make up for the lack of formal qualifications with other skills, including language proficiency.

-Insert Table 1 here-

In Table 2 we restrict the sample to the immigrant population and report summary statistics. The proportion of over-, under-qualified and over-skilled individuals is 22.1%, 9.8% and 17.4%. These figures fall well within the range of estimates reported in the literature, especially those based on Australian data (Green et al., 2007, Mavromaras at al., 2007, Mavromaras and McGuinness, 2012). Relative to matched immigrants, over-qualified workers are less likely to be English proficient, have more years of schooling and arrive to Australia at later ages. They are also more likely to be NESB immigrants, earn higher wages, work in the public sector and in larger firms. By contrast, the under-qualified tend to arrive at earlier ages, are less likely to be NESB immigrants, have more tenure at the firm and are more likely to work in the private sector and in smaller firms. Over-skilled workers are similar to matched workers, except for the fact that they are slightly more prone to be NESB and earn lower wages. We do not find sizable differences among mismatch categories in terms of marital status and hours of work. There are some interesting differences in terms of region of origin. Relative to the other groups, the over-qualified are more likely to come from South and Central Asia, while a significant proportion of the under-qualified (68.5%) comes from Oceania and Northern Europe. The distribution of origins among over-skilled individuals is very similar to the distribution in the total sample.

Finally, although not reported in the table, it is worth noting that only 1 out of 6 over-qualified individuals are also over-skilled or that, alternatively, only 1 out 3 over-skilled individuals are also over-qualified. These figures suggest that worker qualifications may not reflect adequately total work-related human capital.

-Insert Table 2 here-

Table 3 reports results from standard OLS and probit regressions. The probit estimates are marginal effects. We cluster the standard errors at the individual level. English proficiency is not significantly related with the probability of being over-qualified in the linear model. However, it attracts a negative, significant coefficient in the probit model, according to which proficient immigrants are 2.6 pp less likely to be over-qualified than non-proficient immigrants. The estimate is similar when it comes to over-skilling, although in this case the coefficient is

significant only at the 10% level. Finally, language skills are positively and significantly related to the probability of under-qualification. Before discussing to what extent these results hold after controlling for the endogeneity of the language variable, we first examine what are the global determinants of the various types of mismatch.

Years of schooling are a strong predictor of over- and under-qualification. The coefficients are large and have the expected sign. An additional year of schooling raises (decreases) the probability of over-qualification (under-qualification) by between 8.4 and 10.9 pp (5.4 and 8.2 pp). The estimate for over-skilling is lower (-1.9 pp) and its negative sign confirms that education and skills mismatched reflect different phenomena. Over-qualification depends positively on age of arrival, working in the private sector and having Asiatic origins, and negatively on tenure at the job and working in a large firm. Under-qualification and over-skilling are also related to the immigrant's origin, with workers from Southern and Eastern Europe, Oceania and South-East Asia being less prone to be under-qualified than the rest in the linear model, and workers from Southern Europe, Maghreb, Middle East and East Asia being more likely to be over-skilled than the rest in the two models. The prevalence of over-skilling in lower among older individuals and those working more hours. Marital status and previous unemployment experience are not significantly related to any form of mismatch.

Finally, the role of macroeconomic factors is modest, a result that can be explained by the year fixed effects included in the regressions. The year dummies partially factor out yearly fluctuations in the selected aggregate indicators. Still, we find that individual over-skilling is related to the labour market participation rate and the proportion of immigrants in the region.

5. 1. The impact of English proficiency on the probability of mismatch

Next, we switch to the 2SLS and BP estimates. These involve a first stage, English proficiency equation, whose results are reported in Table 4. Since the determinants of English proficiency are very similar across models, we only report results for the over-qualification model.

The 2SLS and BP estimates show that the excluded instrument is highly significant and matches a priori expectations. Among NESB immigrants, late arrival to Australia is negatively correlated with the probability of being English proficient. Beyond age 9, an additional year of delay decreases the likelihood of English proficiency by between 1.3 and 1.5 pp, relative to ESB immigrants. Individuals with more years of schooling, no previous unemployment experiences and working more hours are more likely to be English proficient. This pattern matches a priori expectations, insofar as years of schooling is an efficiency factor and employment and working hours are factors of exposure to language learning. Moreover, workers from Southern and Eastern Europe, Maghreb and Middle East countries, Asia and Latin America are less likely to be proficient.

In the bottom part of the table we report several diagnosis tests for the quality of the instrument. The F-test for the significance of the instrument in the 2SLS model, 92.2, is well above the range of values needed to keep 2SLS bias at low levels (Stock et al., 2002). Similarly, the relative contribution of the instrument to R^2 in the first stage equation, 6.3%, is sizable. We also test for the exogeneity of the instrumented variable by means of the Durbin-Wu-Hausman statistic. The rejection of the null hypothesis suggests that *EP* cannot be regarded as exogenous in the mismatch equation. Similarly, the likelihood-ratio test of the BP model for H_0, $\rho = 0$, signals the presence of endogeneity, implying that the mismatch and English proficiency equations should not be estimated independently.

In Table 5 we report the determinants of mismatch. For reasons of space and noting that the remaining covariates do not change much relative to the OLS and probit estimates, we report only the coefficients of English proficiency. Once the endogeneity of this variable is controlled for, language proficiency emerges as a significant determinant of over-qualification. The results are well above the estimates obtained previously and suggest that assuming exogenous *EP* yields downward-biased predictions. According to the 2SLS model, an immigrant is 35.7 pp less likely to be over-qualified if she is proficient in English. This figure is the LATE, i.e., the average treatment effect on immigrants who are induced into the treatment by the instrumental variable (the complier subpopulation). If we focus, instead, on the effect for an immigrant randomly drawn from the population (ATE) the estimate is somewhat lower, -17.2 pp, and very close to the effect only among those individuals who actually received the treatment (ATET), -18.2 pp.

In contrast, language proficiency raises the probability of under-qualification, although the evidence is less conclusive in this case. The LATE is close to zero and non-significant, while the ATE and ATET are about 8.0 pp. Again, the divergence across models suggests that as far as the impact of English proficiency on mismatch status is concerned, treatment effects are heterogeneous and dependent upon the characteristics of the treated and the compliers. Unfortunately, with non-experimental data it is very difficult to tell who the compliers are. [7]

The fact that English proficiency has a different role when accounting for over- and under-qualification may seem contradicting. However, it should not be so if we consider that

[7] Examining the profile of compliers and non-compliers can be important to understand disparities between local and average effects. However, methods of profiling in the 2SLS context are scarce, complex and limited to binary treatments and instruments. Some examples include Abadie' (2003) local average response function (LARF) and more recently Marbach and Hangartner (2020). Moreover, heterogeneity in treatment responses is driven by both observable and unobservable variables. Finding that compliers and non-compliers are similar in terms of their observable covariates does not imply that we can generalize the LATE to the ATE without invoking additional assumptions.

for some workers language skills are not perfect substitutes of other forms of human capital acquired through schooling. English proficiency raises productivity at the job and thereby increases the range and quality of jobs that immigrants can get. Accordingly, proficient immigrants are less likely to be over-qualified. As far as over-qualification is concerned, formal schooling and host language proficiency are not substitute forms of human capital, for they are associated to opposite effects. Quite the contrary, the results indicate that language skills give immigrants host country-specific human capital that cannot be matched by formal schooling. At the same time, for some workers, this host country-specific human capital makes up for insufficient formal education, enabling them to access jobs for which they are formally under-qualified.

Finally, we find modest evidence that English proficiency reduces the probability of being over-skilled at the job. The 2SLS estimate is very close to the BP coefficients. However, the former is associated with a larger standard error and fails to be statistically significant. [8] The ATE and ATET, about 9.0 pp, are significant only at the 10% level.

6. Sensitivity checks

In this section we examine the robustness of the results to changes in the measurement of mismatches. We also discuss the validity and parametrization of the instrument and examine potential biases arising from panel attrition and transitions in and out of employment.

6.1 Changes in the definition of schooling and skills requirements.

A typical concern in the literature refers to the measurement of educational mismatches. The HILDA dataset does not contain subjective evaluations to assess the gap between the worker's education level and the requirements of the job. This precludes researchers from having a benchmark against which to compare the results from the realized matches approach. Still, we can examine to what extent the estimates are sensitive to the choice of the one standard-deviation range around the mean as required schooling. In Table A2 we report results under two variations of the method. This consists in defining required schooling as either the modal or median level of schooling within the occupation. In the two cases, we obtain a very similar LATE of over-qualification, of about -25.0 pp., a figure that is sensitively below the estimate from the baseline model (-35.7 pp). The differential can be partially accounted for by fact that the baseline estimates correspond to the most stringent classification criteria and, therefore, capture more

[8] This is consistent with Chiburis' (2012) montecarlo simulations of 2SLS and BP models, according to which confidence intervals in the former model tend to be substantially larger than in the later.

salient effects, at least among the subpopulation of compliers. [9] As for the BP models, there are very little variations relative the previous estimates, with average treatment effects in the range of -18 pp when it comes to over-qualification. Likewise, the estimates for under-qualification point to positive and significant effects, comparable in magnitude to those obtained in the baseline regressions.

6.2 Industry-based measures of over- and under-qualification

In the paper, required schooling is calculated at the occupation level. However, mismatches at the industry level may be also informative. For some workers the relevant hierarchy of education qualifications may be based on the distribution within industries. For instance, an increase in the education of workers relative to the industry average may reflect a change in the type of worker hired and thus may reflect a change in required qualifications. Moreover, workers whose education is above the occupation standard may not be over-qualified if in their industry required schooling levels are above average. These workers can be hardly regarded as having "excess education", a perspective that has been highlighted by the few papers have relied on industry cells to measure education mismatches (Liu et al., 2016).

As a robustness check, we combine the 2-digit ANZSCO occupations with 2-digit ANZSIC industry codes, generating a total of 35×28 = 980 detailed industry-occupation cells. Relying on this 4-digit classification reduces the risks of pooling together individuals with different jobs within the same occupation. However, this comes at the cost of having less representative cell sizes and, since we drop groups with fewer than ten observations within each year, ultimately results in a loss of 3,797 obs. relative to the baseline estimates. The results under this 4-digit classification are reported in Table A3. We find little variations relative to the baselines regressions. The local and average impacts of host language proficiency on the probability of over-qualification are again large and significant, about -38.7 pp and -16.0 pp, respectively, while the average effect upon under-qualification rises slightly, from a minimum of 7.5 pp in the baseline model to 9.5 pp under the new categorization.

6.3 Instrument validity

The instrument used in the paper is relevant, for it accounts significantly for differences in language skills between ESB and NESB immigrants. However, instrument validity requires that non-language age-at arrival effects on labour market performance are the same for the two

[9] The prevalence of over-qualification rises from 22.1% in the baseline classification to 35.8% and 36.8% when we use modal and median schooling, respectively. 100% of the workers initially classified as over-qualified are also over-qualified under the other two alternative criteria, whereas two out of five workers who are over-qualified using these alternative thresholds were not initially regarded as over-qualified.

groups of immigrants. As Bleakley and Chin (2010) argue, people coming from poorer countries may face additional barriers to adaptation and these barriers may increase in severity as a function of age at arrival. If, for instance, non-English speaking countries are poorer and have worse education systems, the estimates reported so far may reflect not only differential English-language skills but also differential returns to origin-country schooling. Coming from a country with superior school quality may buffer the potential negative effects that late arrival to the host country may have on the knowledge of the host society, cultural convergence and, ultimately, the probability of having a mismatched job.

To examine this hypothesis, in Table A4 we include interactions between age at arrival and correlates of origin-country school quality as additional controls. We use three different indicators, PISA scores, GDP per capita and the student-teacher ratio in the country of birth. By measuring literacy in different domains (reading, mathematics, and science) among 15-year-old students, average PISA scores provide a direct assessment on the country's school quality. We use data from the most up-to-date PISA survey (2018), which also provides the largest coverage of countries (79). PISA scores are available only recently and cannot be used to proxy school quality at the time of education of most respondents in the sample. Therefore, we must assume that current PISA scores are correlated with school quality in the past. In contrast, GDP per capita and the student-teacher ratio are available for the last 40 years. In this case, we assign each immigrant the corresponding school quality indicator when she was 15 years old. [10] This is a relevant stage, for it provides a snapshot of the country's school quality after completion of primary education.[11]

The principal finding is that the 2SLS and BP estimates for over-qualification remain large and significant. The LATE effect ranges from -28.4 to -49.4 pp depending on the school quality indicator used in the analysis, whereas the ATE and ATET range from -16.4 to -21.0 pp. We also find that the effects of host language proficiency on under-qualification and over-skilling tend to be non-significant in the BP model once we allow for school quality differences. This observation may be indicative of differentials in origin-country school quality that are related to differentials in English-language skills. However, this interpretation has to be taken cautiously, for in some models the loss of significance can be attributed to smaller sample sizes.

In the tables we have also included the interaction term between age at arrival and the school quality indicators to test whether, as hypothesised, the effects of late arrival are buffered by a good education background. The evidence is not very conclusive. In the first two panels of Table A4 we find that the interaction terms are low and mostly non-significant. However, in the

[10] We use longitudinal data from the World Bank's publicly available indicators. Immigrants from countries with missing information were dropped from the regressions.

[11] Other criteria, including 12 and 18 years old and averaging the school quality indicators when the respondent was between 12 and 18 years provided similar regression results.

bottom part of the table the interaction between age at arrival and the student-teacher ratio is negative and well-defined, implying that the negative effects of late arrival to host country on the probability of mismatch are larger among immigrants coming from more crowded education systems.

6.4 Alternative parametrizations of the instrument

We have assumed that immigrants whose first exposure to English was before 9 exhibit similar language skills, regardless of whether they are ESB or NESB immigrants. This assumption seems to be consistent with the pattern of language ability displayed in Figure 2. However, other critical ages are possible, insofar as the language ability gap starts increasing after age 7 and is substantial at older ages. Therefore, it is illustrative to appraise the reliability of the results to changes in the parametrization of the instrument. This is done in Table A5. Setting the critical age at 7 or 11 years yields estimates that come very close to the baseline results. The change is practically negligible in both the 2SLS and the BP models, implying that the point estimates reported in the paper are fairly robust within the range of relevant age thresholds.

6.5 Panel attrition and transitions to and out of employment

The nature of the survey implies that some individuals may not be observed in all years. While the original sample members are augmented with the entrance of new members, there are other individuals that leave the survey for several causes: survey-related reasons (unsuccessful follow-up and refusal) and reasons unrelated to the survey (moves abroad and deaths). Moreover, in our case transitions to and out of employment imply changes in sample composition.

The non-random exit of immigrants from employment and/or the dataset for reasons correlated with mismatch is a potential concern. In our sample, each individual is observed on average for 5.9 out of 14 years. The average entry rate (individuals not present in the sample in the previous period who are currently employed) is 14.6%, and the average exit rate (employed in the previous period and currently not in the sample) is 6.4%. Regressing a dummy equal to 1 for individuals who enter the sample on English proficiency, current mismatch status (yes/no), and year fixed effects, we obtain coefficients equal to -0.008 (p-value = 0.146) for English proficiency, 0.006 (p-value = 0.242) for over-qualification, -0.013 (p-value = 0.118) for under-qualification and 0.030 (t = 0.000) for over-skilling. Therefore, individuals' entry to the estimation sample is mostly uncorrelated with English proficiency and educational mismatches. However, in this parsimonious specification we reject the null that entry is random for over-skilling, implying that selective entry may yield an inconsistent estimate. To check whether this is in fact the case, we add progressively the full set of explanatory variables used in the paper in the entry equation. The resulting coefficients of English proficiency and all mismatch

variables are statistically zero. We also defined a dummy equal to 1 for individuals who left the sample and proceeded likewise, finding that individual attrition is uncorrelated with English proficiency and educational mismatches.

In complementary regressions we focussed on immigrants who are not affected by sample attrition but leave the estimation sample due to changes in their employment status. In this case, the entry and exit rates in the sample are, respectively, 4.4% and 4.9%. Likewise, we considered individuals who enter the sample because they get an employment. The coefficients on English proficiency, over-qualification and overkilling were statistically zero both in the entry and exit equations. In the parsimonious specifications, under-qualification was related to exit from employment in the next period (0.013, p-value = 0.051). However, it was non-significant once we included the full list of controls. This observation gives further evidence that English proficiency and mismatches are orthogonal to sample composition. [12]

7. Conclusions

This paper used micro-data from the 2003-2016 waves of the HILDA dataset to examine to what extent host language proficiency affects the probability of over-qualification, under-qualification and over-skilling among immigrants in Australia. The potential endogeneity of host language proficiency was controlled for by exploiting information on the language spoken during childhood and age at arrival in Australia as a source of external variation. The results, which can be interpreted as causal effects, were based on two complementary econometric approaches, Two Stage Least Squares (2SLS) and a Bivariate Probit (BP). This allowed us to identify and compare local and average treatment effects.

We found that host language proficiency reduces the probability of being over-qualified and, to a lower extent, the probability of over-skilling. In contrast, the likelihood of under-qualification increases with English proficiency. In some cases, average and local estimates differ markedly, a result that suggests that treatment effects cannot be regarded as homogeneous. The results reported in the paper pass well a battery of sensitivity checks, such as accounting for methodological changes in the definition of educational mismatches, the use of very detailed industry-occupation cells when appraising schooling requirements, assessing the validity and testing for alternative parametrizations of the instrument and controlling for attrition and transitions to and out of the labour market.

From a theoretical perspective, the findings of the paper are consistent with the notion that English proficiency raises productivity at the job and thereby increases the range and quality of jobs that immigrants can get. As far as over-qualification is concerned, formal schooling and

[12] The full list of results for entry and exit equations under different combinations of regressors is available from the authors upon request.

host language proficiency are not substitute forms of human capital, for they are associated to opposite effects. Quite the contrary, the results indicate that language skills give immigrants host country-specific human capital that cannot be matched by formal schooling. At the same time, for some workers, this host country-specific human capital makes up for insufficient formal education, enabling them to access jobs for which they are formally under-qualified.

As a related aspect, the paper shows that host language skills are significantly related to educational and skills mismatches. To the extent that mismatches are typically related to lower earnings, it is likely that relevant part of the earnings penalty typically attributed to language shortages in previous works is due to the inability of non-proficient immigrants to access matched jobs. Testing this hypothesis in future research would prove fruitful to understand the channels by which language skills are related to labour market earnings.

From a policy perspective, the results reported here may help policy makers devise strategies and immigration policies that promote and guarantee economic and social stability. In this respect, it would be advisable to provide language courses for immigrants upon arrival. The underutilization of human capital is an issue in both developed and developing economies and a major concern for policy. There are strong grounds for believing that substantial benefits would accrue to individuals, firms and the macroeconomy should policy interventions raise average language skills among immigrants. Moreover, policy responses aimed at reducing over-qualification and over-skilling have tended to focus on fostering labour mobility, matching labour supply with demand and reducing information asymmetries. Without doubting the value of such policy initiatives, the results in this paper put the focus on a novel aspect, language proficiency, that may be of paramount importance among the immigrant population.

References

Abadie, A. (2003). Semiparametric Instrumental Variable Estimation of Treatment Response Models. Journal of Econometrics 113(2): 231–263. https://doi.org/10.1016/S0304-4076(02)00201-4

Akbulut-Yuksel, M., Bleakley, H. and Chin, A. (2011). The effects of English proficiency among childhood immigrants: are Hispanics different? in Latinos and the Economy: Integration and Impact in Schools, Labour Markets, and Beyond, Leal, D., Trejo, S. J. (Eds.). https://doi.org/10.1007/978-1-4419-6682-7_13

Aldashev, A., Gernandt, J., and Thomsen, S. L. (2009). Language usage, participation, employment and earnings: Evidence for foreigners in West Germany with multiple sources of selection. Labour Economics, 16(3): 330-341. https://doi.org/10.1016/j.labeco.2008.11.004

Angrist, J. and Pischke, S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press, 2009.

Baert, S., Cockx, B., Verhaest, D. (2013). Overeducation at the start of the career: steppingstone or trap? Labour Economics, 25: 123–140. https://doi.org/10.1016/j.labeco.2013.04.013

Bauer, T. (2002). Educational mismatch and wages: a panel analysis. Economics of Education Review 21: 221–229. https://doi.org/10.1016/S0272-7757(01)00004-8

Belfield, C. (2010). Over-education: What influence does the workplace have? Economics of Education Review, 29 (2): 236-245. https://doi.org/10.1016/j.econedurev.2009.08.001

Bleakley, H. and A. Chin (2004). Language Skills and Earnings: Evidence from Childhood Immigrants, Review of Economics and Statistics 86: 481–96. https://doi.org/10.1162/003465304323031067

Bleakley, H. and A. Chin (2010). Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants, American Economic Journal: Applied Economics, 2 (1): 165-92. DOI: 10.1257/app.2.1.165

Büchel, F., and Mertens, A. (2004). Overeducation, undereducation, and the theory of career mobility. Applied economics, 36(8). 803-816. https://doi.org/10.1080/0003684042000229532

Budria, S., Colino, A., and Martínez-de-Ibarreta, C. (2019). The impact of host language proficiency on employment outcomes among immigrants in Spain. Empirica, 46: 625–652. https://doi.org/10.1007/s10663-018-9414-x

Chiburis, R., Dasb, J., Lokshin, M. (2012). A practical comparison of the bivariate probit and linear IV estimators, Economics Letters, 117(3): 762-766. https://doi.org/10.1596/1813-9450-5601

Chiswick, B. and Hurst, M. (2000). The Employment, Unemployment and Unemployment Compensation Benefits of Immigrants, in Laurie J. Bassi and Stephen A. Woodbury, eds., Long-Term Unemployment and Reemployment Policies, (Research in Employment Policy, Vol. 2). Stamford, CN, JAI Press, 87-115. https://ssrn.com/abstract=224237

Chiswick, B. and Miller, P. W. (2009). Earnings and Occupational Attainment among Immigrants. Industrial Relations: A Journal of Economy and Society, 48(3): 454-465. https://doi.org/10.1111/j.1468-232X.2009.00568.x

Chiswick, B., Le, A. and Miller, P. (2008). How immigrants fare across the earnings distribution in Australia and the U.S, Industrial and Labour Relations Review, 61 (3): 353-373. https://doi.org/10.1177/001979390806100305

Chiswick, B. and Miller, P. W. (2010). Occupational language requirements and the value of English in the US labour market, Journal of Population Economics, 23, 353-372. https://doi.org/10.1007/s00148-008-0230-7

Cohn, E. and Khan, P. (1995). The wage effects of overschooling revisited. Labour, Economics 2: 67–76. https://doi.org/10.1016/0927-5371(95)80008-L

Dustmann, C., and Van Soest, A. (2001). Language fluency and earnings: Estimation with misclassified language indicators. Review of Economics and Statistics, 83(4): 663-674. https://doi.org/10.1162/003465301753237740

Dustmann, C. and Fabbri, F. (2003). Language proficiency and labour market performance of immigrants in the UK. The Economic Journal, 113: 695-717. https://doi.org/10.1111/1468-0297.t01-1-00151

European Commission (2015). Measuring Skills Mismatch. European Commission Analytical Web Note 7/2015.

Flisi, S., Goglio, V., Meroni, E.C. (2017). Measuring Occupational Mismatch: Overeducation and Overskill in Europe—Evidence from PIAAC, Social Indicators Research, 131: 1211–1249. https://doi.org/10.1007/s11205-016-1292-7

Green, C., Parvinder, K. and Gareth, L. (2007). Immigrant overeducation: Evidence from recent arrivals to Australia, Economics of Education Review, 26(4): 420-432. https://doi.org/10.1016/j.econedurev.2006.02.005

Heckman, J. and Urzua, S. (2010). Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. Journal of Econometrics 156(1): 27–37. https://doi.org/10.1016/j.jeconom.2009.09.006

Hsieh, C.-T., Hurst, E., Jones, C. I., and Klenow, P. J. (2019). The Allocation of Talent and US Economic Growth, forthcoming in Econometrica. https://doi.org/10.3982/ECTA11427

Jovanovic B. (1979). Job Matching and the Theory of Turnover, Journal of Political Economy, 87: 972-990. https://doi.org/10.1086/260808

Leuven E. and Oosterbeek, H. (2011). Overeducation and Mismatch in the Labour Market, in E.A. Hanushek, S. Machin and L. Woessmann, Handbook of the Economics of Education, vol. 4: 283-326. https://doi.org/10.1016/B978-0-444-53444-6.00003-1

Lillard, D.R., Christopoulou, R., Goebel, J., Freidin, S., Lipps, O., Snider, K., KLIPS team. (2016). Codebook for the Cross-National Equivalent File 1980-2015 BHPS – SOEP – HILDA - KLIPS - PSID – SHP - SLID Available at: https://cnef.ehe.osu.edu/data/codebooks/.

Liu, K., Salvanes, K. and Sørensen, E. (2016). Good Skills in Bad Times: Cyclical Skill Mismatch and the Long-Term Effects of Graduating in a Recession. European Economic Review, 84, 2016: 3–17. https://doi.org/10.1016/j.euroecorev.2015.08.015

Marbach, M., and Hangartner, D. (2020). Profiling Compliers and Non-compliers for Instrumental-Variable Analysis. Political Analysis, 1-10. doi:10.1017/pan.2019.48.

Mavromaras, K., McGuinness, S., and Wooden, M. (2007). Over-skilling in the Australian Labour Market, Australian Economic Review, 40 (3): 307-312. https://doi.org/10.1111/j.1467-8462.2007.00468.x

Mavromaras, K. and McGuinness, S. (2012). Overskilling Dynamics and Education Pathways. Economics of Education Review, 31(5): 619–628. https://doi.org/10.1016/j.econedurev.2012.02.006

McGuinness, S. and Bennett, J. (2007). Overeducation and the Graduate Labour Market: A Quantile Regression Approach, Economics of Education Review, 26(5): 521-531. https://doi.org/10.1016/j.econedurev.2005.12.003

McGuinness, S., Bergin, A. and Whelan, A. (2018). Overeducation in Europe: Trends, Convergence and Drivers, Oxford Economic Papers, 70 (4): 994–1015. https://doi.org/10.1093/oep/gpy022

Sattinger M. (1993). Assignment models of the distribution of earnings, Journal of Economic Literature, XXXI: 831–880.

Shields, M. A., and Wheatley Price, S. (2001). Language fluency and immigrant employment prospects: evidence from Britain's ethnic minorities. Applied Economics Letters, 8(11): 741-745. https://doi.org/10.1080/13504850010038678

Spence, M. (1973). Job market signalling. Quarterly Journal of Economics, 87(3): 355-374. https://doi.org/10.1016/B978-0-12-214850-7.50025-5

Stock, J.H., Wright, J. and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. Journal of Business and Economic Statistics, American Statistical Association, vol. 20(4): 518-529. https://doi.org/10.1198/073500102288618658

Summerfield, F and Theodossiou, I (2017). The effects of macroeconomic conditions at graduation on overeducation Economic Inquiry, vol. 55, no. 3: 1370-1387. https://doi.org/10.1111/ecin.12446

Van der Velden, R. and Verhaest, D. (2017). Are Skill Deficits Always Bad? Toward a Learning Perspective on Skill Mismatches, Research in Labour Economics, in: Solomon W. Polachek and Konstantinos Pouliakas and Giovanni Russo and Konstantinos Tatsiramos (ed.). Skill Mismatch in Labour Markets, volume 45: 305-343, Emerald Publishing Ltd.

Verhaest, D. and Omey, E. (2009). Objective over-education and worker well-being: A shadow price approach, Journal of Economic Psychology, 30(3): 469-481.

Verhaest, D. and Omey, E. (2012). Over-Education, Under-education and Earnings: Further evidence on the Importance of Ability and Measurement Bias. Journal of Labour Resources, 33: 76-90.

Verhaest, D. and Van der Velden, R. (2013). Cross-country differences in graduate overeducation, European Sociological Review, 29 (3): 642-653. https://doi.org/10.1016/j.joep.2008.06.003

Yao, Y. and Van Ours, J.C. (2015). Language Skills and Labour Market Performance of Immigrants in the Netherlands, Labour Economics, 34: 76-85. https://doi.org/10.1016/j.labeco.2015.03.005

Zhen, Y. (2013). The effects of English proficiency on earnings of US foreign-born immigrants: Does gender matter. Journal of Finance and Economics, 1(1): 27-41. https://doi: 10.12735/jfe.v1i1p27

Zhu, R. and Chen, L. (2016). Overeducation, over-skilling and mental well-being. The B.E. Journal of Economic Analysis and Policy 16(4). https://doi.org/10.1515/bejeap-2015-0187

Tables

Table 1 – Incidence of mismatch by immigrant status and English proficiency

	Overqualified (%)	Underqualified (%)	Overskilled (%)
Immigrants			
English proficient	20.5	10.8	19.0
	(40.4)	(31.0)	(39.2)
Non English proficient	26.8	6.7	26.4
	(44.3)	(24.9)	(44.1)
Native born	11.1	13.3	19.3
	(31.4)	(34.0)	(39.5)

Notes to Table 1: i) Source: HILDA 2003-2016 waves; ii) standard errors are in parenthesis.

Table 2 – Summary statistics

	All	Matched	Overqualified	Underqualified	Overskilled
Share	**100**	**68.1**	**22.1**	**9.8**	**17.4**
English proficient (%)	75.9	76.5	70.8	83.5	71.1
	(42.8)	(42.4)	(45.5)	(37.1)	(45.3)
Years of schooling	13.4	12.8	16.7	10.1	13.0
	(2.5)	(1.6)	(1.9)	(2.0)	(2.6)
Age	42.6	42.6	42.3	43.8	41.4
	(10.3)	(10.4)	(9.73)	(10.8)	(10.71)
Age at arrival	19.4	18.1	24.1	17.4	19.2
	(12.0)	(11.9)	(11.6)	(10.5)	(11.6)
NESB	38.8	36.0	53.7	24.8	43.7
	(48.7)	(48)	(49.9)	(43.2)	(49.6)
Single (%)	14.3	14.5	14.0	14.2	17.1
	(35.1)	(35.2)	(34.7)	(35.0)	(37.7)
Divorced (%)	6.4	6.6	5.4	7.5	6.6
	(24.5)	(24.8)	(22.6)	(26.3)	(24.9)
Married (%)	79.0	78.7	80.4	77.8	76.2
	(40.7)	(41.0)	(39.7)	(41.6)	(42.6)
Have children (%)	68.8	70.2	63.0	72.4	62.6
	(46.3)	(45.8)	(48.3)	(44.7)	(48.4)
Prior unemployment (last 3 years)	26.9	25.9	31.2	24.2	27.6
	(44.3)	(43.8)	(46.3)	(42.8)	(44.7)
Hourly wage (australian $)	28.7	28.2	32.3	24.1	24.8
	(21.5)	(20.9)	(24.6)	(16.8)	(14.9)
Usual weekly working hours	44.2	44.3	43.7	44.3	41.5
	(10.8)	(10.9)	(10.1)	(11.2)	(10.3)
Tenure (in the actual job) (years)	7.0	7.4	5.4	8.3	6.3
	(7.5)	(7.8)	(5.9)	(8.6)	(7.01)
Private sector (%)	80.7	81.8	74.2	87.3	84.9
	(39.5)	(38.6)	(43.7)	(33.3)	(35.8)
Public sector (%)	15.5	14.7	20.9	9.2	12.7
	(36.2)	(35.4)	(40.7)	(28.9)	(33.3)
Non -profit sector (%)	3.8	3.5	4.9	3.5	2.5
	(19.1)	(18.4)	(21.5)	(18.3)	(15.5)
Number employed at place of work (%)					
Less than 10	32.5	33.0	27.0	41.2	29.2
	(46.8)	(47.0)	(44.4)	(49.2)	(45.5)
Between 10 and 50	23.9	24.4	22.7	23.4	26.8
	(42.7)	(43.0)	(41.9)	(42.3)	(44.3)
Between 50 and 200	20.0	19.9	21.5	17.0	20.0
	(40.1)	(40.1)	(41.1)	(37.6)	(40.2)
More than 200	23.6	22.7	28.9	18.4	24.0
	(42.5)	(41.9)	(45.3)	(38.8)	(42.7)

…..Continues in next page

	All	Matched	Overqualified	Underqualified	Overskilled
Region of origin (%)					
Oceania	19.1	22.0	7.8	24.6	17.5
	(39.3)	(41.5)	(26.8)	(43.1)	(38)
Northern Europe	31.4	32.5	22.5	43.9	28.5
	(46.4)	(46.9)	(41.8)	(49.6)	(45.2)
Southern Europe	5.7	6.5	3.4	5.7	7.2
	(23.2)	(24.6)	(18.1)	(23.1)	(25.9)
Eastern Europe	2.5	2.4	3.8	0.5	2.4
	(15.7)	(15.3)	(19.1)	(7.4)	(15.3)
Maghreb & Middle East	3.8	3.3	5.2	3.7	4.8
	(19.1)	(18)	(22.3)	(19)	(21.4)
South - East Asia	9.7	9.6	12.1	5.7	10.6
	(29.7)	(29.4)	(32.6)	(23.3)	(30.8)
East Asia	5.1	4.2	9.7	0.8	6.4
	(22)	(20.1)	(29.6)	(9.0)	(24.5)
South & Central Asia	10.4	7.3	22.8	3.4	11.3
	(30.5)	(26.1)	(42)	(18.1)	(31.6)
N. America	2.7	2.7	3.0	1.6	2.3
	(16.2)	(16.3)	(17.1)	(12.7)	(14.9)
Latin - America	2.8	2.5	4.4	1.3	3.5
	(16.5)	(15.6)	(20.5)	(11.2)	(18.3)
Subsaharian Africa	6.6	6.8	5.3	8.2	5.5
	(24.8)	(25.1)	(22.4)	(27.5)	(22.8)
Australian State (%)					
New south Wales	34.5	33.2	39.7	32.2	34.3
	(47.5)	(47.1)	(48.9)	(46.7)	(47.5)
Victoria	23.7	23.1	25.7	23.0	27.3
	(42.5)	(42.2)	(43.7)	(42.1)	(44.5)
Queensland	16.8	18.5	11.3	18.1	13.9
	(37.4)	(38.8)	(31.6)	(38.5)	(34.6)
South Australia	6.8	7.3	5.9	5.2	8.3
	(25.2)	(26.1)	(23.6)	(22.2)	(27.5)
West Australia	12.3	12.2	11.8	14.6	12.4
	(32.9)	(32.7)	(32.3)	(35.3)	(32.9)
Tasmania	1.1	1.2	0.8	1.3	0.9
	(10.4)	(10.8)	(8.7)	(11.2)	(9.6)
Northern Territory	1.5	1.6	0.5	2.4	0.9
	(12.0)	(12.7)	(7.2)	(15.2)	(9.6)
Australia Capital Territory	3.2	2.8	4.3	3.3	2.1
	(17.6)	(16.6)	(20.3)	(17.8)	(14.2)

Notes to Table 2: i) Source: HILDA 2003-2016 waves; ii) standard errors are in parenthesis.

Table 3 – OLS and probit estimates

	OLS			PROBIT		
	Overqualified	Underqualified	Overskilled	Overqualified	Underqualified	Overskilled
English language proficiency	**-0.022**	**0.034** ***	**-0.028** *	**-0.026** **	**0.029** **	**-0.025** *
	(0.015)	**(0.012)**	**(0.017)**	**(0.011)**	**(0.011)**	**(0.015)**
Years of schooling	0.109 ***	-0.054 ***	-0.019 ***	0.084 ***	-0.082 ***	-0.019 ***
	(0.005)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)
Age	0.000	0.001	-0.002 ***	0.000	0.001	-0.002 **
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Max (0; age at arrival - 9)	0.003 ***	0.001	0.001 **	0.001 ***	0.000	0.001 *
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Married (*base category single)*	-0.013	0.005	-0.003	-0.015	0.014	-0.002
	(0.018)	(0.019)	(0.022)	(0.015)	(0.012)	(0.021)
Divorced	0.010	-0.013	0.019	0.016	0.010	0.023
	(0.024)	(0.027)	(0.031)	(0.021)	(0.017)	(0.029)
Have children (yes/no)	-0.003	-0.020	-0.033 *	-0.006	-0.028 **	-0.034 **
	(0.016)	(0.017)	(0.018)	(0.013)	(0.012)	(0.017)
Prior unemployment (last 3 years)	0.010	0.002	0.000	0.010	-0.004	-0.002
	(0.009)	(0.009)	(0.013)	(0.009)	(0.007)	(0.012)
Ln (Usual weekly working hours)	-0.018	0.010	-0.227 ***	-0.014	0.012	-0.216 ***
	(0.019)	(0.019)	(0.024)	(0.018)	(0.014)	(0.023)
Ln (Tenure in the actual job)	-0.011 ***	0.001	-0.006	-0.009 ***	0.001	-0.005
	(0.003)	(0.004)	(0.004)	(0.003)	(0.002)	(0.004)
Private sector (base: public sector)	0.046 **	-0.020	0.030	0.038 ***	-0.017	0.032
	(0.021)	(0.016)	(0.021)	(0.015)	(0.014)	(0.021)
Non -profit sector	-0.035	0.047	-0.046 *	-0.027	0.071 **	-0.060 *
	(0.034)	(0.035)	(0.026)	(0.023)	(0.034)	(0.033)
Number employed at place of work (base category: between 10 and 50)						
Less than 10	0.015	0.015	-0.024	0.020 *	0.004	-0.023
	(0.014)	(0.013)	(0.016)	(0.012)	(0.009)	(0.016)
Between 50 and 200	-0.018	0.008	-0.011	-0.016	0.008	-0.010
	(0.013)	(0.013)	(0.016)	(0.011)	(0.011)	(0.016)
More than 200	-0.038 **	0.033 **	0.004	-0.046 ***	0.010	0.005
	(0.016)	(0.013)	(0.019)	(0.013)	(0.011)	(0.018)
Region of origin (base: N. Europe)						
Oceania	-0.019	-0.042 **	-0.012	0.002	-0.013	-0.011
	(0.018)	(0.018)	(0.018)	(0.019)	(0.013)	(0.021)
Southern Europe	0.013	-0.055 **	0.076 **	0.017	-0.021	0.073 **
	(0.024)	(0.025)	(0.038)	(0.026)	(0.017)	(0.032)
Eastern Europe	0.052	-0.053 **	-0.011	0.052	-0.059	-0.007
	(0.056)	(0.025)	(0.045)	(0.032)	(0.036)	(0.046)
Maghreb & Middle East	0.049	0.004	0.095 **	0.039	-0.008	0.083 **
	(0.046)	(0.031)	(0.048)	(0.031)	(0.021)	(0.039)
South - East Asia	0.072 **	-0.053 ***	0.017	0.067 ***	-0.024	0.020
	(0.034)	(0.018)	(0.032)	(0.018)	(0.018)	(0.032)
East Asia	0.043	-0.017	0.089 ***	0.038 *	-0.028	0.086 ***
	(0.036)	(0.022)	(0.034)	(0.023)	(0.026)	(0.03)
South & Central Asia	0.088 ***	-0.001	0.038	0.049 ***	-0.014	0.039
	(0.029)	(0.017)	(0.028)	(0.019)	(0.023)	(0.027)
N. America	-0.043	-0.027	0.004	-0.001	-0.005	0.004
	(0.047)	(0.029)	(0.047)	(0.044)	(0.028)	(0.053)
Latin - America	0.066	-0.031	0.049	0.055 *	-0.027	0.046
	(0.044)	(0.032)	(0.048)	(0.031)	(0.027)	(0.043)
Subsaharian Africa	-0.032	0.003	-0.028	0.005	0.042 **	-0.027
	(0.028)	(0.032)	(0.027)	(0.024)	(0.019)	(0.029)

…..Continues in next page

Macroeconomic controls (by year - state)						
% of female workers/ total workers	0.014	-0.002	0.018	0.016	-0.002	0.020
	(0.011)	(0.011)	(0.017)	(0.011)	(0.009)	(0.018)
Participation rate	0.008	0.004	-0.017 **	0.007	0.002	-0.020 **
	(0.007)	(0.006)	(0.009)	(0.007)	(0.005)	(0.009)
Share part-time workers	0.002	-0.001	-0.006	0.002	-0.002	-0.005
	(0.006)	(0.006)	(0.008)	(0.006)	(0.004)	(0.009)
Unemployment rate	0.005	0.009	-0.009	0.001	0.006	-0.013
	(0.009)	(0.009)	(0.013)	(0.009)	(0.007)	(0.014)
GDP per capita (x1000)	0.001	-0.003	-0.002	0.001	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)
GDP yearly growth rate	0.000	0.000	0.001	0.001	0.000	0.001
	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.004)
% of immigrant people in region	-0.015	-0.007	0.026 *	-0.011	0.001	0.030 *
	(0.011)	(0.011)	(0.016)	(0.011)	(0.008)	(0.016)
% of population with university degree	0.010	-0.005	0.002	0.006	-0.007	0.001
	(0.007)	(0.006)	(0.009)	(0.007)	(0.005)	(0.009)
Constant	-2.320 ***	1.032	1.310			
	(0.811)	(0.797)	(1.075)			
Australian regions fixed effects	yes	yes	yes	yes	yes	yes
Time fixed effects	yes	yes	yes	yes	yes	yes
R^2	0.49	0.21	0.06			
F statistic	8.44	6.39	5.74			
	(p = 0.000)	(p = 0.000)	(p = 0.000)			
Chi 2 statistic				1125.4	477.6	284.9
				(p = 0.000)	(p = 0.000)	(p = 0.000)
No. of observations	11,224	11,224	11,224	11,224	11,224	11,224

Notes to Table 3: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

Table 4 – Determinants of English proficiency

	2SLS - First stage	BP
Max (0; age at arrival - 9) × NESB	**-0.015** ***	**-0.013** ***
	(0.002)	**(0.001)**
Years of schooling	0.016 ***	0.014 ***
	(0.003)	(0.003)
Age	-0.001	0.000
	(0.001)	(0.001)
Max (0; age at arrival - 9)	0.004 ***	0.006 ***
	(0.001)	(0.001)
Married (*base category single)*	-0.031	-0.035 *
	(0.021)	(0.021)
Divorced	-0.016	-0.014
	(0.026)	(0.031)
Have children	0.001	0.000
	(0.018)	(0.018)
Prior unemployment (last 3 years)	-0.033 ***	-0.029 ***
	(0.012)	(0.011)
Ln (Usual weekly working hours)	0.046 *	0.050 **
	(0.024)	(0.023)
Ln (Tenure in the actual job)	-0.002	-0.002
	(0.004)	(0.004)
Private sector (base: public sector)	0.000	-0.005
	(0.026)	(0.024)
Non -profit sector	0.000	0.000
	(0.033)	(0.034)
Number employed at place of work (base category: between 10 and 50)		
Less than 10	-0.001	0.000
	(0.016)	(0.016)
Between 50 and 200	0.023	0.025
	(0.016)	(0.016)
More than 200	-0.019	-0.013
	(0.018)	(0.018)
Region of origin (base: N. Europe)		
Oceania	0.026	0.050
	(0.018)	(0.031)
Southern Europe	-0.308 ***	-0.215 ***
	(0.045)	(0.029)
Eastern Europe	-0.375 ***	-0.261 ***
	(0.053)	(0.033)
Maghreb & Middle East	-0.214 ***	-0.159 ***
	(0.061)	(0.038)
South - East Asia	-0.320 ***	-0.225 ***
	(0.041)	(0.026)
East Asia	-0.323 ***	-0.228 ***
	(0.051)	(0.033)
South & Central Asia	-0.182 ***	-0.165 ***
	(0.041)	(0.031)
N. America	0.036 **	0.236 ***
	(0.018)	(0.085)
Latin - America	-0.261 ***	-0.186 ***
	(0.075)	(0.047)
Subsaharian Africa	0.006	0.013
	(0.023)	(0.039)

…..Continues in next page

Macroeconomic controls	yes	yes
Australian regions fixed effects	yes	yes
Time fixed effects	yes	yes
R^2 first stage	0.311	
Partial R^2	0.063	
F statistic	92.21 (p = 0.000)	
Durbin -Wu Hausman test	17.06 (p = 0.000)	
Chi 2 statistic		2103.53 (p = 0.000)
Log likelihood		-6,808.680
Rho		0.601
Likelihood ratio test of ρ = 0		14.967 (p = 0.000)
No. of observations	11,224	11,224

Notes to Table 4: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

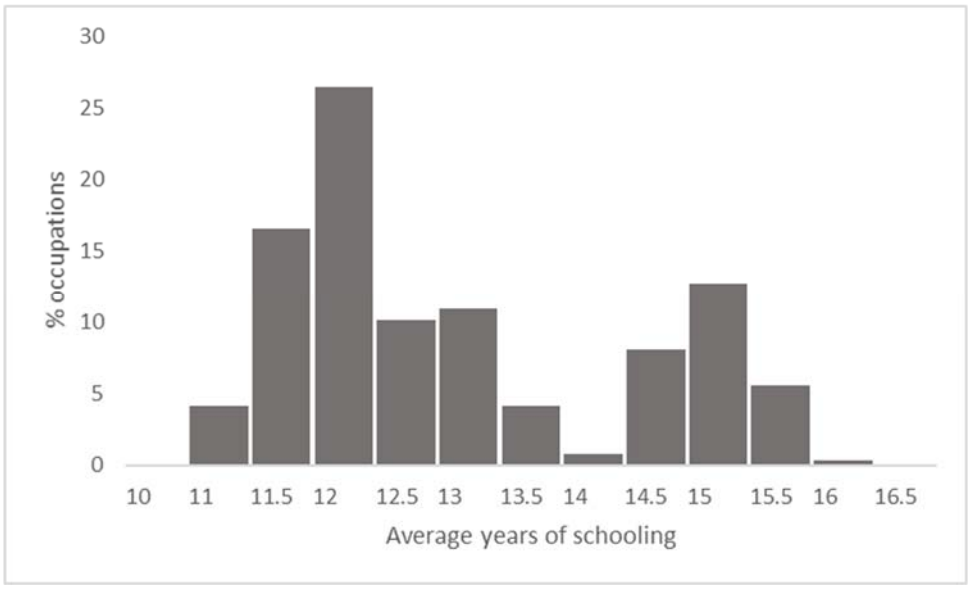
Table 5 – Mismatch and language proficiency

	2SLS			BP		
	Overqualified	Underqualified	Overskilled	Overqualified	Underqualified	Overskilled
English language proficiency						
LATE	-0.357 ***	-0.001	-0.094			
	(0.087)	(0.051)	(0.085)			
ATE				-0.172 ***	0.086 ***	-0.089 *
				(0.017)	(0.025)	(0.047)
ATET				-0.182 ***	0.075 ***	-0.090 *
				(0.019)	(0.016)	(0.048)
No. of observations	11,224	11,224	11,224	11,224	11,224	11,224

Notes to Table 5: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

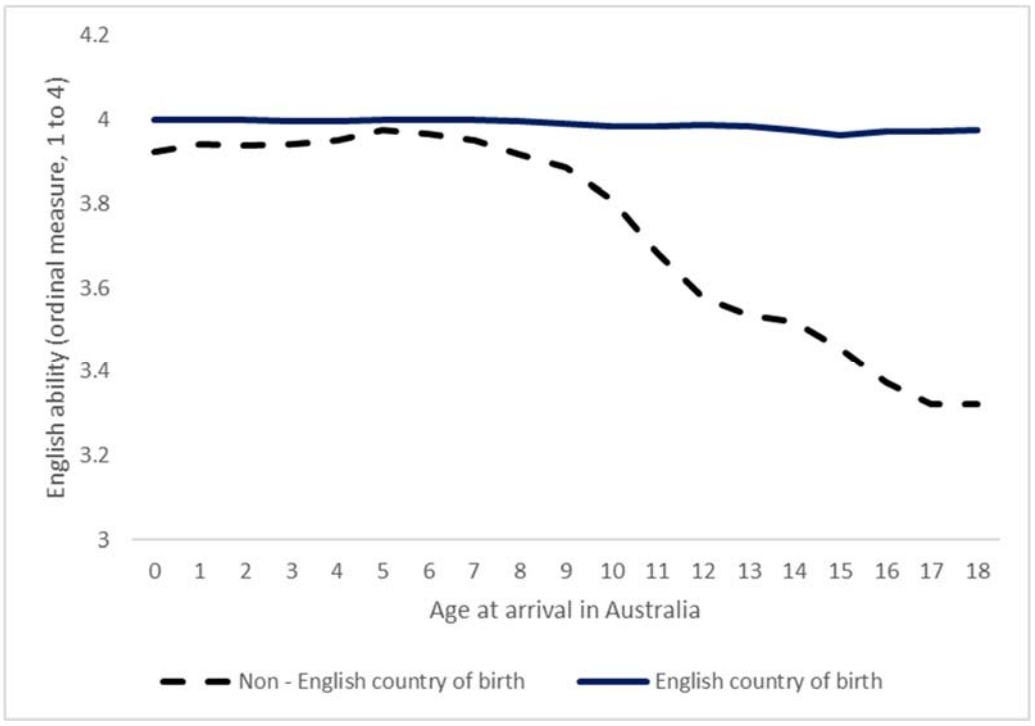
Figures

Figure 1. Frequency distribution of average schooling across occupations



Notes to Figure 1: i) Source: HILDA 2003-2016 waves.

Figure 2. English proficiency among English-speaking background (ESB) and non-English-speaking background (NESB) immigrants, by age at arrival.



Notes to Figure 2: i) Source: HILDA 2003-2016 waves.

Appendix

Table A1. Immigrants by country of birth

Panel A. English speaking countries

Rank by N	Country	N	% of group
1	United Kingdom	2,883	42.00%
2	New Zealand	1672	24.36%
3	India	582	8.48%
4	South Africa	434	6.32%
5	Fiji	254	3.70%
6	United States of A	174	2.53%
7	Ireland	156	2.27%
8	Canada	129	1.88%
9	Papua New Guine	124	1.81%
10	Zimbabwe	79	1.15%
11	Singapore	77	1.12%
12	Zambia	54	0.79%
13	Malta	50	0.73%
14	Mauritius	48	0.70%
15	Tanzania	30	0.44%
16	Tonga	26	0.38%
17	Solomon Islands	23	0.34%
18	Kenya	22	0.32%
19	Jamaica	14	0.20%
20	Ghana	9	0.13%
21	Seychelles	6	0.09%
22	Bahamas	6	0.09%
23	Malawi	3	0.04%
24	Vanuatu	3	0.04%
25	Botswana	2	0.03%
26	Nigeria	2	0.03%
27	Trinidad and Toba	2	0.03%
28	Marshall Islands	1	0.01%
Total English speaking countries		6,865	100.00%

Panel B. Non - English speaking countries

Rank by N	Country	N	% of group
1	Vietnam	335	7.67%
2	Philippines	333	7.63%
3	China (exclu	314	7.19%
4	Sri Lanka	252	5.77%
5	Netherlands	156	3.57%
6	Hong Kong	149	3.41%
7	Germany	143	3.28%
8	Malaysia	133	3.05%
9	Italy	125	2.86%
10	Nepal	120	2.75%
11	Bangladesh	108	2.47%
12	Poland	104	2.38%
13	Indonesia	102	2.34%
14	Lebanon	92	2.11%
15	Croatia	91	2.08%
16	Colombia	82	1.88%
17	Iran	75	1.72%
18	Chile	72	1.65%
19	France	67	1.53%
20	Egypt	67	1.53%
21	Russian Fed	64	1.47%
22	Spain	63	1.44%
23	North Mace	63	1.44%
24	Romania	61	1.40%
25	Pakistan	57	1.31%
26	Turkey	55	1.26%
27	Greece	52	1.19%
28	Yugoslavia,F	52	1.19%
29	Czech Repul	49	1.12%
30	Iraq	49	1.12%
	Subtotal, top 30 countries	3,485	79.93%
	Subtotal, other (59) countries	875	20.07%
	Total Non - Engish speaking countries obs.	4,359	100.00%

Notes to Table A1: i) Source: HILDA 2003-2016 waves.

Table A2 – The effects of English proficiency on education and skills mismatches – Alternative schooling thresholds

	2SLS		BP	
	Overqualified	Underqualified	Overqualified	Underqualified
Required schooling: modal				
English language proficiency				
LATE	-0.252 ***	0.026		
	(0.089)	(0.074)		
ATE			-0.180 ***	0.071 ***
			(0.026)	(0.016)
ATET			-0.186 ***	0.069 ***
			(0.028)	(0.014)
No. of observations	11,224	11,224	11,224	11,224
Required schooling: median				
English language proficiency				
LATE	-0.247 ***	0.049		
	(0.086)	(0.075)		
ATE			-0.186 ***	0.087 ***
			(0.038)	(0.019)
ATET			-0.191 ***	0.083 ***
			(0.041)	(0.017)
No. of observations	11,224	11,224	11,224	11,224

Notes to Table A2: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

Table A3. The effects of English proficiency on education and skills mismatches – 4-digit industry-occupation cells

	2SLS		BP	
	Overqualified	Underqualified	Overqualified	Underqualified
English language proficiency				
LATE	-0.387 ***	0.082		
	(0.099)	(0.061)		
ATE			-0.158 ***	0.117 ***
			(0.027)	(0.01)
ATET			-0.166 ***	0.095 ***
			(0.03)	(0.006)
No. of observations	7,427	7,427	7,427	7,427

Notes to Table A3: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

Table A4 – The effects of English proficiency on education and skills mismatches - Adding school quality controls

	2SLS			BP		
	Overqualified	Underqualified	Overskilled	Overqualified	Underqualified	Overskilled
PISA 2018 results						
English language proficiency						
LATE	-0.284 ***	-0.073	-0.025			
	(0.073)	(0.054)	(0.074)			
ATE				-0.184 ***	-0.005	-0.045
				(0.029)	(0.039)	(0.046)
ATET				-0.197 ***	-0.006	-0.044
				(0.032)	(0.042)	(0.046)
Pisa 2018 × Max(0, age at arrival - 9)×1000	-0.006	0.023	0.034	0.000	0.000	0.035
	(0.019)	(0.014)	(0.019)	(0.000)	(0.000)	(0.019)
No. of observations	7,886	7,886	7,886	7,886	7,886	7,886
GDP per capita when aged 15						
English language proficiency						
LATE	-0.346 ***	-0.017	-0.083			
	(0.100)	(0.055)	(0.098)			
ATE				-0.164 ***	0.061	-0.064
				(0.024)	(0.040)	(0.048)
ATET				-0.170 ***	0.060	-0.064
				(0.026)	(0.037)	(0.050)
Ln(GDP) × Max(0, age at arrival - 9)×1000	-0.001 *	0.000	0.000	0.000 *	0.000	0.000
	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
No. of observations	10,236	10,236	10,236	10,236	10,236	10,236
Pupil-teacher ratio primary when aged 15						
English language proficiency						
LATE	-0.494 ***	-0.057	-0.005			
	(0.127)	(0.068)	(0.106)			
ATE				-0.199 ***	0.073	0.032
				(0.027)	(0.107)	(0.056)
ATET				-0.210 ***	0.073	0.031
				(0.031)	(0.107)	(0.052)
P-t ratio × Max(0, age at arrival - 9)×1000	0.346 ***	0.132 *	0.264 *	0.115	0.021	0.259 ***
	(0.102)	(0.077)	(0.088)	(0.073)	(0.091)	(0.086)
No. of observations	6,419	6,419	6,419	6,419	6,419	6,419

Notes to Table A4: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.

Table A5 - The effects of English proficiency on education and skills mismatches – Alternative instrument parametrizations

	Instrument: Max(0, age at arrival - 7) × I(NESB inmigrant)					
	2SLS			BP		
	Overqualified	Underqualified	Overskilled	Overqualified	Underqualified	Overskilled
English language proficiency						
LATE	-0.342 ***	-0.001	-0.087			
	(0.084)	(0.050)	(0.082)			
ATE				-0.172 ***	0.081 ***	-0.084 *
				(0.017)	(0.020)	(0.045)
ATET				-0.182 ***	0.072 ***	-0.084 *
				(0.019)	(0.013)	(0.048)
No. of observations	11,224	11,224	11,224	11,224	11,224	11,224
	Instrument: Max(0, age at arrival - 11) × I(NESB inmigrant)					
LATE	-0.376 ***	-0.001	-0.100			
	(0.091)	(0.053)	(0.088)			
ATE				-0.176 ***	0.091 ***	-0.095 *
				(0.020)	(0.031)	(0.053)
ATET				-0.186 ***	0.078 ***	-0.096 *
				(0.022)	(0.021)	(0.056)
No. of observations	11,224	11,224	11,224	11,224	11,224	11,224

Notes to Table A5: i) Source: HILDA 2003-2016 waves; ii) *** denotes significant at the 1% level, ** denotes significant at the 5% level; * denotes significant at the 10% level; iii) standard errors (in parenthesis) are clustered at the individual level.