Heriot-Watt University

Accountancy, Economics, and Finance Working Papers

Working Paper 2024-12

Skill vs Education Types of Labour Mismatch and Their Association with Earnings

Vsevolod Iakovlev

December 2024

Keywords: Earnings, Education, Skill, Labour Mismatch

JEL: D31, I24, J24

Skill vs Education Types of Labour Mismatch and Their Association with Earnings

Vsevolod Iakovlev*
Heriot-Watt University

Working Paper
December 16, 2024

Abstract

This work aims to determine whether the difference between skill and education-based measures of labour mismatch affects the estimates of the labour mismatch-earnings relationship as specified by Verdugo and Verdugo's (1989) version of over, required and under-education (ORU) Mincer earnings function. The analysis employs crossectional data for 26 countries from the 1st Cycle of the OECD Survey of Adult Skills (PIAAC) conducted between 2011 and 2012. The preliminary results of the graphical analysis show that education and skill mismatch may exhibit opposite relationships with earnings at the country level. Specifically, over and under-education are found to be positively associated with median earnings, whereas over and under-skilling show a negative association. To investigate the source of the opposite correlations, an error components model is used. Additionally, the paper explores the heterogeneity in earnings and labour mismatch across a set of commonly used controls. The analysis produces mixed coefficient estimates for under and over-education but predominantly negative estimates for under and over-skilling at both individual and market levels. The market-level unobserved heterogeneity is found to be driving the coefficients away from zero. Although removing it often leads to a slight loss of magnitude, some exceptions exhibit a change of sign or loss of statistical significance. It is, thus, concluded that education and skill mismatch should be distinguished both conceptually and empirically and, if used as a proxy for each other, are unlikely to produce accurate results in the analysis of the Mincer earnings function.

JEL Classification: D31, I24, J24
Keywords: Earnings, Education, Skill, Labour Mismatch

*Email: vsevolod.v.iakovlev@gmail.com

Contents

1 Introduction

Since Mincer's "Investment in human capital and personal income distribution" (1958), the relationship between education and earnings has been of interest to labour economists for over six decades. One of the directions the research has taken focuses on the concept of labour mismatch, which refers to the mismatch between the characteristics of a worker defining their competencies and corresponding requirements of their job, and its effect on earnings. One of the central issues for the analysis of this relationship is the measurement of mismatch, which is particularly important when the labour mismatch analysis aims to draw conclusions about workers' skills. Due to the lack of datasets containing direct measures of workers' skills, the traditional approach would involve using education as a proxy for skill. This has led to a concern among the researchers that imposing an equivalence between educational and skill mismatch may lead to invalid empirical results (Allen & Van der Velden, 2001). As the datasets containing direct measures of skills became more available, this concern gained popularity among the researchers, some of whom, e.g. Allen et al. (2013) and Pellizzari and Fichen (2017), suggested alternative skill mismatch measures. However, due to the latent nature of labour mismatch, the advantage of the new skill-based measures over the education-based ones is not obvious. Furthermore, one can argue that such labour market theories as human capital, job competition and labour market friction models provide theoretical justification for the skill-educational mismatch equivalence. This imposes a question of whether the use of an alternative measurement framework leads to different results for the analysis of the effect of labour mismatch on earnings.

This paper attempts to analyse the implications of various labour mismatch measures on the results of a labour mismatch specification of the Mincer (1974) earnings function. More specifically, the effect of the difference between education and skill-based measures of mismatch on the analysis of earnings is of particular interest to this work. The paper starts by providing background for the research question and reviewing related literature in Section 2, which is followed by an outline of the major educational and skill mismatch measurement frameworks in Section 3 and Section 4 discussing the main labour market theories' interpretations of the relationship between skill and education and their linkage with earnings. Section 5 describes data employed for the empirical analysis. This is followed by the cross-country comparison of the mismatch shares in Section 6.1, analysis of the correlation coefficients in Section 6.2, and evaluation of the mismatch measures' contribution to the out-of-sample prediction performances of the Mincer equation in Section 6.3. Section 6.4 presents an overview of the earnings and market-level mismatch shares distributions across gender, age, migration, education and skill groups. Section 7 describes the econometric model, whose estimation results are analysed in Section 8. Finally, Section 9 concludes the analysis and outlines the directions for further research.

The main results of the paper suggest that education and skill mismatch should be distinguished both conceptually and empirically and, if used as a proxy for each other, are unlikely to produce accurate results in the analysis of the Mincer earnings function. Moreover, the coefficient estimates for under and over-matching may be sensitive to the choice of a mismatch measure. Additionally, considerable heterogeneity is observed in the educational and skill mismatch across gender, age and migration status and requires a separate investigation.

2 Background and related literature

In the late 1960s, economists' attention was drawn to education during their search for the reasons behind income inequality. This was preceded by theories connecting the distribution of earnings with "ability" and "chance". Mincer (1958) argues that neither of the two proposed factors has led to a significant contribution to our understanding of the causes of income inequality. The former appears to be challenging to find an empirically appropriate measure for. For instance, an intelligence quotient is not a good fit because its normal distribution contrasts with the largely skewed distribution of earnings. The latter is not helpful either because its stochastic nature prevents the researchers from drawing economic intuition. Building upon the work of Friedman (1953) that establishes a link between rational choice and personal income distribution, Mincer (1958) suggests a model with a focus on the investment in training while assuming identically distributed "ability" and "chance". According to the model, occupations require different amounts of training while undertaking which the individuals receive no income. Therefore, a worker is entitled to compensation determined by the present values of their lifetime earnings at the time when an occupation is chosen. Among other implications of the model, Mincer shows a positive relationship between one's training and earnings. It is not, however, until his 1974 model that we see the classic Mincer equation:

$$\ln w = a_0 + \rho s + \beta_0 x + \beta_1 x^2 + \varepsilon, \tag{1}$$

that establishes a relationship between the natural logarithm of earnings w on the LHS and schooling s, linear and quadratic experience terms x and x^2 on the RHS.[2]

The empirical support of some theoretical implications of the Mincer equation was at the centre of discussion among labour economists for a few decades. For instance, using 1940-1950 data, Heckman et al. (2003) show that, as Mincer's model predicts, no amount of experience can result in higher earnings for a worker with less schooling than for a worker with more schooling. However, they reject the hypothesis using 1960-

[2]For the derivation of equation (1), as well as the discussion and empirical analysis of the two models' implications, see work by Heckman et al. (2003).

1970 data and find convergence in the log-earnings-experience profiles for some schooling levels using 1980-1990 data. Similarly, Rumberger (1987) shows that the nature of the seemingly straightforward relation between schooling and earnings is unapparent when the former fails to match the job's requirements. His work summarises the debate on the validity of the human capital model in the context of schooling surplus. This view relies on the premise of firms maximising the use of workers' skills by adapting to the changes in prices and production technology through the substitution of inputs. Thus, earnings are associated with workers' productivity, which is argued to be linked with schooling via skill. Therefore, a surplus in schooling should still have a positive effect on earnings because the firms employ and fully utilise workers as long as their marginal product is above the wage. This, however, is not the case in the job competition model developed by Thurow (1975), where firms hire workers based on estimated costs of their training, which are predicted with observable characteristics, such as education. Since the firms focus on solving the cost minimisation problem, they may decide not to take advantage of the schooling surplus. Hence, workers' return to schooling surplus can potentially be zero or negative. Rumberger (1987) approached the controversy by modifying equation (1) to what later became commonly known as over, required and under-education[3] (ORU) Mincer equation:

$$\ln w = a_0 + \rho_o s_o + \rho_r s_r + \rho_u s_u + \boldsymbol{\beta}\mathbf{X} + \varepsilon, \tag{2}$$

where over- and under-education are respectively defined by the required level of schooling s_r and individual level of schooling s_i as

$$s_o = \begin{cases} s_i - s_r & \text{if } s_i > s_r \\ 0 & \text{if } s_i \leq s_r \end{cases} \text{ and } s_u = \begin{cases} s_r - s_i & \text{if } s_i < s_r \\ 0 & \text{if } s_i \geq s_r \end{cases}.$$

Thus, Rumberger suggests that if the human capital theory is incorrect, we should expect $\widehat{\rho}_o \leq 0$ and $\widehat{\rho}_r > \widehat{\rho}$. This specification typically has a lesser focus on experience, which is now buried in the vector of covariates $\mathbf{X}$ along with other personal characteristics.

The ORU specification has been used by various papers to produce estimates of the educational mismatch effect on earnings, which do not lack consistency. Hartog (2000) reviews 45 sets of results for five countries, different years covering the period between 1969 and 1992, and three main measurement frameworks: job analysis, realised matches, and direct self-assessment. The works reviewed seem to agree that (i) $\widehat{\rho}_r > \widehat{\rho}_o$, (ii) $\widehat{\rho}_o > 0$ but $\widehat{\rho}_o < \widehat{\rho}_r$, and (iii) $\widehat{\rho}_u < 0$. Hartog (2000) finds that (ii) is the only result that holds for all works, whereas (i) and (ii) feature exceptions for some years and countries. Nevertheless, all three points hold despite various measurement frameworks. This seems to be an important result, suggesting the robustness of the conclusions above and the

[3]Note that original Rumberger's (1987) specification does not feature under-education.

indifference of the ORU specification to the means of measurement. However, the review amends the results of the models utilising alternative inputs. Verdugo and Verdugo (1989) were the first to approach educational mismatch, not in terms of the gap in years of schooling but in terms of classification. They modify equation (2) by replacing s_o and s_u with binary variables for over- and under-education:

$$\ln w = a_0 + \rho_o overed + \rho_u undered + \boldsymbol{\beta}\mathbf{X} + \varepsilon, \tag{3}$$

and compute results suggesting the opposite of (ii) and (iii).[4] Their work started a controversy over the validity of such an approach. The results were challenged by Cohn (1992), suggesting misspecification of the model, and Gill and Solberg (1992) questioning the empirical strategy. Verdugo and Verdugo addressed the criticism in their 1992 reply. They refuse the idea that their results are driven by the comparison of individuals from low and high-paid jobs by pointing out that the analysis involves multiple occupation groups with similar levels of earnings. Furthermore, they defend the use of binary over and under-education variables alongside years of schooling by arguing that highly educated workers are not necessarily expected to be productive at their jobs, thereby undermining the assumption that total return to education is larger than occupation-specific ones. However, it did not solve the controversy. Hartog (2000) lists other works that use Verdugo and Verdugo's specification and find similar results but still omit it as a model producing stand-apart results. Nevertheless, since the respondents can be classified into one of the three categories using the output of any mismatch measure, equation (3) is preferred for the analysis involving multiple different measurement frameworks to ensure their compatibility.

Another measurement issue, which is of special interest to this paper, comes from the distinction between education and skill. Allen and Van der Velden (2001) analyse the approach of assignment theory to educational mismatch. This view has several assumptions that vary across different models, but common features include an explicitly formulated assignment problem, which links the personal characteristics of workers with earnings (Sattinger, 1993). Allen and Van der Velden argue that assignment theory implies educational mismatch to be both the necessary and sufficient condition for skill mismatch and vice versa.[5] They find results suggesting that skill mismatch among Dutch university graduates does not account for a significant proportion of the educational mismatch effect on earnings, which contradicts the predictions derived from the assignment theory. Although their results reveal that over-skilling has a negative effect on wages of its own, the magnitude appears to be small.

[4] Verdugo and Verdugo (1992) also suggest a version of the model with continuous variables for over- and under-education.

[5] It is unclear which model Allen and Van der Velden (2001) refer to, however, their description of assignment theory is mostly similar to the differential rents model developed by Ricardo (1951).

Their work has started a new brunch of literature dedicated to the distinction between skill and educational mismatch, which they draw by analysing workers' responses to two survey questions aiming to reveal underutilisation/deficit of skills.[6] A similar approach is taken by Di Pietro and Urwin (2006). Although their results contradict the ones of Allen and Van der Velden (2001) in the context of on-the-job search, they also find no decrease in the wage penalties associated with over-education of Italian university graduates when accounting for over-skilling. Di Pietro and Urwin interpret these results as opposing the assignment theory and suggest that the discrepancy between skill and education could be potentially explained by a simple variation in ability, which is only weakly related to earnings. This conclusion, however, does not align with the findings of Green and Zhu (2010). Building upon the earlier work by Green and McIntosh (2007), they combine the data from multiple UK surveys to identify the effects of "real" and "formal" over-qualification, where the difference between the two concepts refers to the presence of skill underutilisation among the over-educated workers. They find that there's a steeper increase in the pay penalties associated with real rather than formal over-qualification, which supports the prediction of assignment theory. Thus, although there is little disagreement about the importance of differentiating between skills and education, the extent to which they account for the wage penalties associated with labour mismatch is unclear.

Leuven and Oosterbeek (2011) discuss a few identification problems that could lead to the inconsistency of the empirical results, including unobserved heterogeneity caused by relying solely on educational mismatch and measurement error that is common for the self-reported measures of skill mismatch. Both issues could potentially be solved by applying the instrumental variable methodology – the main well-documented solution to unobserved heterogeneity. However, this approach had limited success due to the absence of credible instruments. Furthermore, Leuven and Oosterbeek (2011) argue that the existing measures of skill mismatch suffer from the lack of appropriate theoretical basis relying instead on the data-driven approach.[7] This implies that acquiring a reliable measure of skills on its own would fail to solve the problem. Specifically, simply comparing one's skills to the occupation-specific average with some statistically determined cutoff implicitly assumes that all workers fully deploy their skill endowment regardless of their match quality.

To address this, Pellizzari and Fichen (2017) suggests a measurement framework of skill mismatch based on a formal theory, which is outlined in Section 6. An explicit economic model is what differentiates the Pellizzari and Fichen (2017) framework from the

[6] This measure is often referred to as direct self-assessment (DSA), see Section 6 for the details.

[7] It is worth mentioning that some studies form a separate branch of the literature by relying on the surveys that are compatible with the databases containing skill requirements to construct a skill mismatch measure. For instance, Lindenlaub (2017), Lise and Postel-Vinay (2020) and Guvenen et al. (2020) use O*NET descriptors aggregated using Principal Component Analysis (PCA).

alternative measures that utilise direct data on skills, such as one developed by Desjardins and Rubenson (2011), who compare a worker's position in the distribution of skill to their position in the distribution of the corresponding task engagement scores. The empirical application of the model comes down to comparing a worker's skill to the occupation-specific critical values determined by the 5th and 95h percentiles of the skill distribution of the well-matched workers, who are identified using a self-reported measure. In the early stage of its development, this approach was criticised by Allen et al. (2013), stating that tying skill requirements to the occupation groups reduces heterogeneity on the demand side while the heterogeneity on the supply side is unchanged. They argue that it leads to a strong covariance between workers' skills and the corresponding requirements, resulting in controversial outcomes. Instead, Allen et al. (2013) construct their own measure that is conceptually similar to the one of Desjardins and Rubenson (2011) but relies on the difference between the standardised variables for skill and relevant task engagement. Recent applied literature suggests that both types of skill mismatch measures are used by the researchers. For instance, by applying Pellizzari and Fichen (2017) model to the first round of PIAAC data (OECD, 2012), Pivovarova and Powers (2022) find that the workers who have migrated to the USA are more likely to be over-matched especially in the early years of their life in the country, which reduces their wages and standards of living. Alternatively, by utilising Allen et al. (2013) measure along with a repeated cross-section consisting of the 1994 IALS, 2003 ALL and 2012 PIAAC surveys, Shin and Bills (2021) obtain the results suggesting that the USA skill mismatch is mainly determined by the job-related variables rather than personal characteristics, including gender and migration status. Although their work is based on a wider range of datasets, the difference in the result might be potentially attributed to the alternative skill mismatch measures, which demonstrates the importance of the methodology choice even in the context of a single type of labour mismatch.

Such discrepancy between the mismatch measures, which partly motivates this work, has prompted multiple methodological papers comparing the construct and output of the measures in various contexts. Flisi et al. (2014) provide an extensive review covering 20 specifications of 3 education and 4 skill-based measures applied to PIAAC data for 17 European countries. The bulk of their analysis focuses on a narrower list of 5 specifications, which are selected by the Principle Component Analysis (PCA) and combined in a single measure, as well as their socio-economic determinants such as country of residence, education level, gender, age and migration. They find that women are more likely to be over-skilled and simultaneously (in both skill and education) mismatched but spot no differences in over-education across genders. The older workers are reported to have a higher chance of both over-education and over-skilling compared to the middle-aged workers, while the younger ones only exhibit a higher chance of being over-skilled. Fi-

nally, their most intuitive result suggests that higher-educated workers are more likely to suffer from all types of over-matching, including the simultaneous one.

This paper adopts the agenda set by Flisi et al. (2014) while taking a different turn on the context and methodology applied to the comparison of the skill and education-based measures of labour mismatch. For this purpose, the output of 3 education-based and 3 skill-based measures with multiple specifications is compared across various socio-economic characteristics. However, rather than combining the multiple frameworks in a single compound measure, the analysis is conducted separately for each of the measures. This approach allows to preserve of the special features of each framework, which are unique in their definitions of labour mismatch and could be capturing different aspects of the matching outcomes. Furthermore, the specifications are selected on the basis of both their predictive power and distinct structure as opposed to the sole ability to explain the variation in the data. Another major difference is the context. Instead of seeking to estimate average predicted probabilities of mismatch, this work focuses on investigating the explanatory power of the mismatch measures in the Mincer earnings function. This, to a certain degree, has been attempted by Desjardins and Rubenson (2011). While citing a variety of education-based measures and the extensive literature on using them as predictors in the Mincer equation, they decide to focus solely on the skill mismatch, which is derived using a predecessor measure of Allen et al. (2013) developed by Krahn and Lowe (1998). They find that being over-skilled (high skill scores & low skill engagement) in literacy on average results in a 4% wage penalty, whereas those who are under-skilled (low skill scores & high skill engagement) tend to earn 21% higher wages. Since, in both cases, the mismatched workers are only compared to the low-skilled, well-matched workers rather than to the combined pool of the well-skilled workers, these estimates provide limited insight into the association between skill mismatch and earnings. This paper improves on their methodology by using more recently developed frameworks of Pellizzari and Fichen (2017) and Allen et al. (2013) that represent the two main alternative approaches to measuring skill mismatch featured in the literature. Finally, rather than computing country-specific estimates as reported in Flisi et al. (2014) or adding a set of corresponding dummies to the statistical model as done by Desjardins and Rubenson (2011), this paper exploits the variation between the countries by setting up a market-level fixed effects model with the markets defined as country-specific industries. Not only does this approach allow us to account for the unobserved heterogeneity, but it also estimates the market-level variation, which, as we find, is an important component of the association between labour mismatch and earnings.

3 Labour mismatch measurement frameworks

As we saw in Section 2, much of the disagreement about the effect of mismatch on earnings was caused by the differences in the measurement methodologies. The literature suggests a wide range of mismatch measurement frameworks, which could be classified based on a variety of characteristics.

Firstly, given the interest of the paper, it is reasonable to split the measures based on the nature of the input variable: a measure of skill or education. Educational mismatch measures have some advantages when compared to skill mismatch ones. The measures of education are well standardised and have clear intuition, e.g. years of schooling or International Standard Classification of Education (ISCED), whereas skill measures are unique to the dataset of choice and require a theoretical framework to enable economic interpretation. This imposes a challenge for the comparison of results within the field because it is unclear whether potential variation originates from the features of a chosen survey or a genuine effect that was not captured before. Another issue is that unlike education, which can be interpreted as an investment, signal, etc., a skill measure does not necessarily show the level of skill that the workers deploy in their occupations, which undermines its linkage with earnings.

Another common distinction is between objective and subjective measures. The latter relies on the genuineness and precision of the respondents, which makes it vulnerable to measurement error. Such measures include Direct and Indirect Self-Assessment (DSA and ISA). DSA is obtained by asking the respondent to evaluate their level of skill or education compared to their job's day-to-day demands. Verhaest and Omey (2006) distinguish between the measures based on the questions regarding the interviewee's education and skill deployment. Traditionally, the reviews of mismatch measures, such as one conducted by Flisi et al. (2017), classify DSA as an educational mismatch measure. Nevertheless, considering its similarity with some self-reported skill mismatch measures, such as one utilised by Allen and Van der Velden (2001), and that it refers to job performance rather than formal requirements, the two versions of DSA can be regarded as a skill-mismatch measure. ISA, on the other hand, is based on the responses regarding the workers' attained and required levels of education. Verhaest and Omey (2006) split ISA measures also based on the questions referring to the education required to perform the job as opposed to the one required to get the job. It could be argued that the former refers to skill and the latter to qualification. However, both questions refer to the ability of the respondents' education to equip them with the necessary skills to perform their jobs, which the second question additionally puts in the context of a specific firm's recruitment process. Although the two could potentially have different implications in some labour market theories, e.g. human capital and signalling, they both refer to education in its qualification sense and, therefore, are measures of educational mismatch.

The two popular objective educational mismatch measures that do not rely on the respondents' self-assessment are job analysis (JA) and realised matches (RM), first used by Rumberger (1987) and Verdugo and Verdugo (1989), respectively. JA derives mismatch level by comparing the attained education of the workers with the required education for their occupation group. The latter is usually defined by an occupation classification, e.g. International Standard Classification of Occupations (ISCO). This measure is conveniently straightforward and regarded by some as "conceptually superior" (Hartog, 2000). Nevertheless, in practice, it is sometimes a challenge to find an appropriate classification for the dataset of interest. Additionally, such generalised and rarely updated classifications as ISCO may not be capturing special features of a particular labour market or a specific period. Unlike JA, RM defines the education required based on the distribution feature in the data. In this framework, a worker is said to be over(under)-education if their attained education is more than one standard deviation above(below) the modal or mean education for their occupation group. One of the main points of criticism of RM is the arbitrariness of the classification threshold. Although it is common to use one standard deviation, some works feature alternative cutoffs, such as 0.5 standard deviations that Tsay et al. (2005) justify with an underestimation concern. Another RM parameter is the measure of central tendency. Some works, including the original paper by Verdugo and Verdugo (1989), give preference to the mean. However, since the distribution of education often features multiple peaks, the mode is arguably more appropriate. Furthermore, its lower sensitivity to outliers makes it likely to produce more accurate results than the mean (Kiker et al., 1997).

Before moving on to an objective measure of skill mismatch, one should note the conceptual difficulty of designing one. It has been mentioned that the measures of skill lack clear economic interpretation. DSA avoids this issue by estimating skill mismatch without preliminary measurement of the skill itself. This, however, only allows producing a subjective measure. An objective measure requires a theoretical framework that would link a measure of skill with the concept of skill mismatch. An example of such a framework is a model developed by Pellizzari and Fichen (2017) (PF). Using original notation, suppose every worker i is characterized by a skill endowment η_i and some level of skill s_i, deployed at job j, such that the following utility function is maximised:

$$U_{ij} = w_{ij} - 1[y_{ij} < 0]F - c_i(s_i), \tag{4}$$

where $w_{ij} = \gamma_i y_{ij}$ denotes wage, which is proportional to the output of the match y_{ij} (i.e. $\gamma_i \geq 0$), $F \geq 0$ is a cost associated with producing negative output, and $c_i(s_i)$ is the cost function of deploying skill s_i that takes the value of δs_i (with $\delta \geq 0$) if the level of deployed skill exceeds endowment η_i and zero otherwise. Finally, the output of the match is a function of deployed skill with a locally constant marginal product that decreases

after some threshold, fixed operational cost k_j, and returns to deployed skill β_j, i.e.

$$y_{ij} = \begin{cases} \beta_j s_i - k_j & \text{if } s_i \leq max_j \\ \beta_j max_j - k_j & \text{if } s_i > max_j, \end{cases} \tag{5}$$

where min_j and max_j are critical points in the skills distribution defined by the values of s that result in zero and maximum output, respectively. Pellizzari and Fichen's model then uses the two critical values above to define skill mismatch. A worker is called over(under)-skilled if their skill endowment is above(below) $max_j(min_j)$, and well-skilled if $min_j \leq \eta_i \leq max_j$. It can be shown that optimal skill deployment values are $s^*_{\text{well-skilled}} = \eta_i$, $s^*_{\text{under-skilled}} = \min_j$ and $s^*_{\text{over-skilled}} \in [max_j, \eta_i]$. Empirical identification of min_j and max_j comes from the respondents' answers to the two interview questions aiming to reveal (i) the ability to do a more demanding job with the current skill endowment and (ii) the need for additional training to do the current job. Workers who answer negatively to both questions are assumed to be well-skilled. Since the optimal skill deployment level for well-skilled workers is simply their endowment, we can derive min_j and max_j based on their distribution of η_i within each occupation group. Over and under-skilled workers are then classified based on the derived values of min_j and max_j. Although Pellizzari and Fichen's framework does rely on DSA in identifying the well-matched workers, it classifies the rest of the workers in a distribution-driven manner that addresses the skill deployment issue mentioned at the beginning of the section. Most importantly, it links a measure of skill with earnings.

Although Pellizzari and Fichen (2017) were the ones to develop the economic model above, the measure itself was first featured in a report by OECD (2013). Shortly after Allen et al. (2013) criticise the framework suggesting that by computing the classification thresholds within the occupation groups, PF reduces heterogeneity on the demand side while making no similar adjustment on the supply side, thereby ignoring the potentially strong correlation between the skill scores and requirements within the occupation groups and leading to paradoxical results. To address this concern, they suggest an alternative measure which relies on the data for skill engagement (skill use). Specifically, their measure involves computing z-scores of the skill and engagement variables. If there are multiple variables available describing skill engagement, their average is used instead. Skill engagement z-scores are then subtracted from the ones of the skill variable. A worker is classified as over-skilled (under-utilised) if the difference between the z-scores exceeds 1.5 and under-skilled (over-utilised) if the value is below -1.5. One can spot a similarity between this measure and the one developed by Krahn and Lowe (1998), which classifies workers into four groups defined by their position in the skill scores and skill engagement distribution: high-skilled match (high score & high engagement), low-skilled match (low score & low engagement), high-skilled mismatch (high score & low

engagement), and low-skilled mismatch (low score & high engagement).[8] This suggests that skill engagement-based measures are another established class of skill mismatch measures along with the skill requirement-based (OECD, 2013, Pellizzari and Fichen, 2017) and self-reported measures (Allen and Van der Velden, 2001, Di Pietro and Urwin, 2006, Green and Zhu, 2010).

The last feature that divides mismatch measures, but manifestly not the least one, is the unit of measure. In the educational mismatch context, this distinction goes back to the 1992 argument between Verdugo and Verdugo suggesting to measure mismatch using binary variables for over- and under-education on the one side and Cohn, Gill and Solberg defending the traditional measurement in the years of over-, under- and required schooling on the other. The two approaches have been reported to produce the results leading to opposite conclusions (Hartog, 2000). In the context of skill mismatch, this distinction is more relevant for the unit of the input variable, which in the case of DSA would be binary but could also be continuous and unique if a dataset features a measure of skill.

In summary, the literature contains a variety of labour mismatch measurement frameworks that could be differentiated in multiple ways. Furthermore, such measures as RM are determined by the parameters that affect the resulting classification. This raises two questions: Which of the measures ought to be used for the prediction of earnings, and whether rarely feasible and complicated skill-based measures lead to results that are different from the ones of education-based measures?

4 Skill vs education in labour market theories

It could be argued that the reason behind using educational mismatch as a proxy for skill mismatch is purely practical – limited availability of the direct skill data. However, employing this approximation for the empirical analysis still requires a theoretical justification, which is often obtained from certain labour market theories' interpretations of the relationship between education and skill as well as their linkage with earnings. Therefore, the validity of the empirical approximation has implications for the validity of the theories. Hence, it is useful to outline the frameworks that are supported and the ones doubted by the results of this work. Leuven and Oosterbeek (2011) provide a brief overview of the role of educational mismatch in six different labour market theories. This section discusses the same list of views complemented with the assignment theory (Sattinger, 1993) to offer a summary of various interpretations of the relationship between skill and education in the context of earnings. The objective of the section is to

[8] Due to the more statistically robust nature of Allen et al. (2013) measure, the alternative (Krahn & Lowe, 1998) is not included in the analysis.

identify the theories that differentiate between skill and education little enough to make the assumption of equivalence between the two concepts appropriate for applied research.

Human capital theory (Becker, 1964) considers education as an investment, which is paid off by an increase in worker's productivity that leads to higher earnings. Productivity and earnings have a direct link that originates from neoclassical economic theory, suggesting that profit-maximising firms hire workers as long as their marginal product is above the costs of acquiring it. The earnings are thus determined by the workers' productivity. Additionally, it is assumed that the firms are able to adapt to changes in prices and production technologies by substituting inputs, so the positive relationship between productivity and earnings is robust to over-education. The link between education and productivity, however, is not direct but goes through skill.[9] In the human capital view, education is just one of the factors contributing to a worker's skill. However, it assumes that the positive relationship between the two is significant enough to produce the returns that partially explain income inequality. This is where the human capital model brings education and skill close to equivalence. In practice, it enables the analysis of the Mincer function without accounting for skill, which effectively imposes the equivalence.[10]

Human capital theory associates marginal products with workers, which contrasts with the job competition model (Thurow, 1975) that associates marginal products with jobs (Rumberger, 1987). In this model, firms fill in the positions while aiming to minimise the costs of training the workers for the jobs. The costs of training are determined by the gap between the job requirements and the skill level of the workers. Since the skills are assumed to be unobservable, the firms use education to predict the training costs and make the hiring decisions. Like human capital theory, the job competition model also admits that skill and education are different concepts but assumes that one is a strong predictor of the other. This has similar implications for the formulation of a statistical model using the Mincer equation.

The relationship between education and skill in the assignment theory is less obvious. Sattinger (1993) suggests that there are three main types of assignment models. These feature slightly different approaches to the two concepts of interest. The one that arguably contrasts with the human capital and job competition models the most is the linear programming optimal assignment model (Koopmans & Beckmann, 1957). It features no hierarchical assumption with regard to either jobs or workers. Although both workers and "machines" differ in certain characteristics, there is no link connecting the two with earnings. Instead, the earnings are determined by the outputs associated with alternative assignments. It could be consistent with Koopmans and Beckmann's model to view skill and education as two of the worker characteristics, the relationship between which is

[9] Rumberger (1987) gives an overview of explanations for the education-skill-productivity relationship.

[10] Although the Mincer function usually includes terms for experience, it serves as another factor affecting skill alongside education rather than as a measure of skill.

ambiguous, but neither is linked with earnings in the Mincerian sense because the effect of alternative values of skill and education on the outcome of the assignment problem is not obvious. The differential rents model (Ricardo, 1951), however, does allow for a hierarchical assignment, in which higher-skilled workers tend to be assigned to more demanding jobs. In Sattinger's formulation, each worker is solely characterised by their skill, although it is suggested that such concepts as education or ability could be used instead. This does not necessarily imply the equivalence but only that the formulation could potentially be extended to an array of characteristics affecting the productivity of the worker-job match. Finally, Roy's (1951) sectoral model combines the features of the two assignment models above. Formally, it can be written as a special case of the linear programming optimal assignment problem, and like the differential rents model, it is solved by the decisions of profit/earning-maximising firms and workers. Unlike the models above, Roy's model is formulated in terms of occupations rather than jobs (e.g. using the original formulation, catching rabbits or fishing for trout). Hence, the earnings are determined by how many projects each worker is able to complete (rabbits or trout to catch), given their choice of occupation. Nevertheless, the role of skill or a skill measure is similar. Rather than being a factor in the earnings function, it illustrates a natural or gained inclination of a worker for a certain occupation, which affects the assignment problem. Although the assignment theory does not allocate distinct roles to education and skill (like the human capital or job competition models, where one is a predictor for the other), it does not make an explicit assumption of the equivalence between the two concepts, as argued by Allen and Van der Velden (2001), but focuses on the solution of the assignment problem as the main determinant of the earnings distribution.

Spence's (1973) signalling model provides education and skill with the most distinct roles so far. Like in human capital theory, education is an investment. However, due to asymmetric information, the return comes not from a positive effect on skill and productivity but from a costlier signal of the worker's innate skill. This relies on the premise that a naturally high-skilled worker would have an easier time pursuing a certain level of education than a naturally low-skilled one. Thus, there exists an education level that would be unprofitable for a low-skilled worker to obtain because the earnings compensation they receive for being recognised as a high-skilled one is below the costs of the education level. Intuitively, Spence's model implies a positive relationship between skill and education because higher skill makes a higher education level less costly to attain. However, this is much further from the equivalence compared to the human capital and job competition models because, in certain cases, the workers are incentivised to fake their skills via educational choices. Additionally, the signalling model has an interesting implication for the educational mismatch discussion. The level of educational mismatch in the economy is dependent upon the definition of the required education. The job

requirements, as defined by the firms, are likely to be distribution-based and depend on the signalling value of education given the distribution of skill across workers. The education level that is actually required to perform a job is zero, which implies that any positive investment in education leads to over-education.

Another theory that contributes to educational mismatch literature is Sicherman and Galor's (1990) model of career mobility, which focuses on the career paths of the workers. In this view, the workers aim to maximise their expected lifetime earnings by allocating finite time among various jobs. Earnings are defined as a function of human capital, which in turn is an increasing function of education and innate skill. This creates two potential sources of returns to education: a direct one via an increase in human capital and an indirect one via a career path improvement. A curious implication of over-education in this formulation is that a worker may choose a job with lower requirements than their attained education if it yields a higher probability of promotion. The career mobility model suggests a relationship between education and skill that incorporates features from human capital, signalling and assignment theories. More specifically, education is considered an investment in human capital, but skill is exogenously fixed, and there exists no direct linkage between the two. Arguably, Sicherman and Galor take education and skill as far from equivalence as possible.

Some economists explain the phenomena of over-education by a macroeconomic concept of labour market frictions. These usually refer to imperfect information and insurance markets, heterogeneity, slow mobility, labour market capacity, etc. and are typically modelled with a matching function taking the number of unemployed workers and open vacancies as inputs (Pissarides, 2000). These models often define earnings with a wage bargaining equation, which is based on the Nash bargaining approach and does not feature either skill or education. Leuven and Oosterbeek (2011) reviews some of the works that incorporate the concept of skill into the search and matching model. The models tend to have two equilibria: one where no worker chooses to work in a job with lower requirements than their attained education and one where some workers choose to work in a job with lower requirements, which is argued to be determined by the gap between productivity levels and skill distribution across workers (Albrecht & Vroman, 2002), productivity and quit rate (Gautier, 2002), and a possibility to pursue a job with higher return (Dolado et al., 2009). However, none of the authors above draws a distinction between skill and education and uses the two concepts interchangeably, which imposes the equivalence.

Finally, it is worth making a note of the role of preferences in educational mismatch literature. Intuitively, the workers who receive higher utility from the process of acquiring an education are more likely to end up over-educated and vice versa. This introduces an additional source of return to education through a utility function, which complicates the

relationship between earnings and educational decisions. Oosterbeek and Van Ophem (2000) conduct an empirical analysis using a Cobb-Douglas utility function with the net present value of lifetime earnings and schooling as inputs. They find that the marginal rate of return to schooling is higher than one to earnings, suggesting that the standard Mincerian model underestimates returns to schooling. Additionally, they find that social background and "innate ability" have a positive relationship with preferences for schooling and a negative one with the marginal rate of return. The relationship between skill and education in the context of preferences and earnings is similar to the one in the career mobility model. The difference originates from the sources of returns to education. In Sicherman and Galor's (1990) model, the indirect effect of education on the returns comes from an improvement in the career path. In contrast, the preferences theory suggests a return in the form of positive utility gained from the process of schooling, which is not related to earnings.

In summary, the literature features a wide range of interpretations of the skill-education relationship in the context of earnings. The theories of human capital, job competition, labour market frictions, and, from Allen and Van der Velden's (2001) point of view, assignment models facilitate the use of data on education for the analysis of skill mismatch. It is challenging, however, to impose a similar empirical framework using signalling, career mobility, preferences and, arguably, assignment frameworks.

5 Survey of Adult Skills (PIAAC)

The data used for the analysis is provided by the First Cycle of the OECD Survey of Adult Skills (PIAAC) conducted between 2011 and 2012 in 35 countries, 26 of which are included in the analysis. The key selling point of PIAAC data is an assessment of three fundamental abilities of the respondents to evaluate and engage with texts, interpret and communicate numerical information, and perform practical tasks using digital communication tools, which are mapped on a 500-point scale. The three sets of skills are referred to as literacy, numeracy and problem-solving, respectively. In addition, the survey also collects data on demographic characteristics, education and training, social and linguistic background, employment status and income, as well as the use of cognitive, interaction, social, physical and learning skills. This section starts with a brief note on motivation for using the PIAAC dataset for this work, which is followed by a description of the data cleaning process, variable creation and summary statistics.

The PIAAC dataset has recently become a common choice for the analysis of labour mismatch in the contexts of productivity, earnings, horizontal mismatch, job tasks and cognitive skills, university enrolment rates, etc. To give a few examples, McGowan and Andrews (2015) find that over-skilling and under-qualification are the main drivers of

Table 1: Summary statistics: earnings and education

	Earnings	Highest qual.	Obtained ISCO SL	Required ISCO SL	Obtained years of ed	Required year of ed.
N	72563	68888	68888	72563	71854	70996
mean	14.60	8.02	2.67	2.57	13.14	12.55
std	10.20	3.84	0.90	0.98	2.97	3.34
min	1.71	1	1	1	3	3
25%	6.97	5	2	2	11	11
50%	11.96	6	2	2	13	12
75%	19.63	12	4	4	15	15
max	66.77	16	4	4	23	23

Notes: Earnings – hourly earnings including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

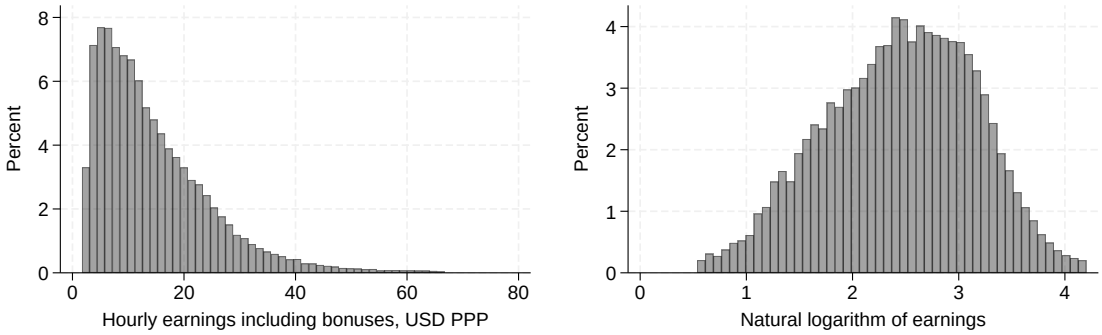
the negative relationship between workers' productivity and labour mismatch. They extend their results using the Second Cycle of the survey in their follow-up work (2017), suggesting that better allocation of skills is associated with higher productivity. Nieto and Ramos (2017) show that the well-documented wage penalty associated with over-education is partially explained by the lower skill level of the over-educated workers compared to the equally educated but employed in more demanding jobs. Montt (2017) evaluates the implications of horizontal in addition to vertical mismatch and finds that working outside one's field does not lead to a wage penalty unless a horizontally mismatched worker is also over-educated. Pouliakas and Russo (2015) use skill mismatch to compute cognitive skill demand to analyse its relationship with task complexity, which they find to be significant. Finally, Castro et al. (2022) evaluates the effect of a shock in tertiary education participation on the magnitude of over-education and over-skilling. Their results suggest that the impact on skill mismatch was similar to the one on education mismatch, although this conclusion only seems to hold for LAC countries, where over-skilling estimates are lower than OECD ones. In summary, PIAAC data is commonly used to research various topics related to labour mismatch, which makes it a good choice for the methodological analysis of the measurement frameworks.

The variables that are essential to the analysis include country of residence, employment status, hourly earnings including bonuses (PPP corrected USD) and current job occupation group in 1-digit International Standard Classification of Occupations 2008 (ISCO). Data instances containing missing values for any of the variables listed above are dropped from the dataset. Furthermore, any respondents who are unemployed or out of the labour force are omitted from the analysis. Other variables used in the analysis are not essential and may contain missing values.

Table 1 contains summary statistics for the earnings and education variables. As expected, the distribution of hourly earnings is notably skewed to the right (see Figure 1). The highest qualification is a categorical variable mapped on a 1 to 16 scale, where

a

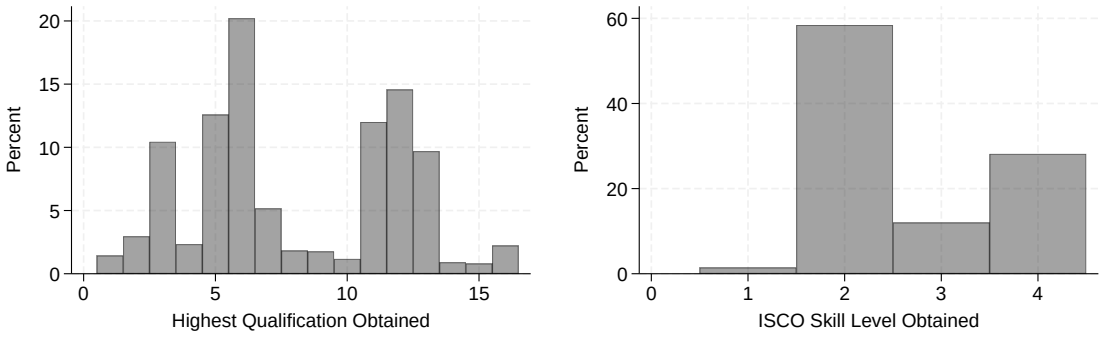
Figure 1: Hourly earnings including bonuses (USD PPP) and its natural logarithm



Notes: The distribution of hourly earnings is trimmed at the 1st and 99th percentiles to avoid outliers.

1 is no qualification or below International Standard Classification of Education 1997 (ISCED) level 1 and 14 is ISCED level 6 (doctoral decree).[11] Obtained ISCO skill level[12] has 4 categories and is derived from the highest qualification, using a mapping provided by the International Labour Organization (2012). Required ISCO skill level is derived by mapping the respondents' ISCO occupation groups on the same 1 to 4 scale (ILO, 2012). Finally, the last two variables in Table 1 refer to years spent in full-time education and years of education required to get a respondent's current job, respectively. The distributions of obtained and required education are similar for both sets of variables, with the mean values being slightly higher for the obtained education.

Figure 2: Obtained qualification and ISCO skill level



One potential piece of criticism may come from shrinking the obtained highest qualification variable with 16 categories to ISCO skill level with 4 categories. This may raise a concern that ISCO skill level is a poor approximation for the distribution of qualification levels. However, it can be seen from Figure 2 that the majority of observations are

[11]Highest qualification values of 15 and 16 are not ordered and correspond to a foreign qualification and an unidentified higher education degree, respectively.

[12]Note that ISCO skill level is derived from an education variable and is not a measure of skill.

Table 2: Summary statistics: skills

	Not challenged	Need training	Literacy	Numeracy	Problem-solving
count	72563	72563	72561	72561	51224
mean	0.84	0.36	271.88	268.40	279.61
std	0.37	0.48	46.01	50.03	41.79
min	0	0	66.39	24.85	8.04
25%	1	0	243.75	238.32	252.10
50%	1	0	276.45	272.73	282.43
75%	1	1	304.57	303.51	309.25
max	1	1	410.65	430.98	480.01

Notes: Not challenged – answered positively to "Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job?" Need training – answered positively to "Do you feel that you need further training in order to cope well with your present duties?"

clustered around several main qualification values: 3 (ISCED level 2), 5-7 (ISCED level 3), and 11-13 (ISCED level 5), which correspond to high school dropouts, those finished high school, and university graduates, respectively. According to the ISCO mapping, the respondents with the highest qualification of 1 are assigned skill level 1, those with qualifications between 2 and 10 are assigned skill level 2, a qualification value of 11 corresponds to skill level 3, and 14 to 16 are equivalent to skill level 4. Given this mapping and the distributions presented in Figure 2, one can argue that the ISCO skill level is an appropriate approximation for the highest qualification.

Figure 3: Skill scores

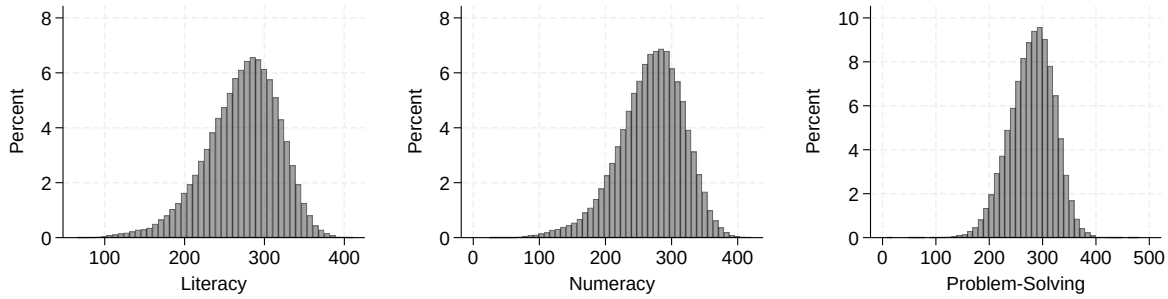


Table 2 contains summary statistics for the variables used to measure skill mismatch. The first two are binary variables indicating whether the respondents feel that they have the skills to cope with more demanding duties and whether they feel the need for further training to cope well with the present duties, respectively. The table shows that 84% are not challenged enough, and 36% need additional training. Although these two variables are not explicit measures of skill, they are used to compute DSA and the Pellizzari-Fichen frameworks. The second three variables contain the results of the PIAAC assessment of the respondents' literacy, numeracy and problem-solving skills, respectively, mapped on a 500-point scale. The variables are computed by averaging 10

Table 3: Frequencies and averages: ISCO occupation groups

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
HS managers	3592	0.05	22.43	12.0	294.71	298.06	294.73	0.40	43.78
Professionals	14882	0.21	18.22	12.0	294.43	293.58	292.51	0.62	40.60
Techc-s & assoc.	10691	0.15	14.99	11.0	283.73	282.65	286.77	0.53	40.26
LS managers	554	0.01	14.17	7.0	280.62	277.81	287.16	0.50	39.67
Clerical support	7778	0.11	12.56	6.0	278.51	273.06	282.62	0.69	39.18
Craft & related	7727	0.11	10.06	5.0	255.74	254.85	264.90	0.13	38.38
Operat. & assem.	5915	0.08	9.32	5.0	253.75	251.10	259.49	0.19	40.46
Service and sales	13697	0.19	9.24	6.0	264.17	256.28	271.97	0.69	36.70
Element. occup.	7141	0.10	7.86	5.0	239.44	230.84	257.78	0.55	39.89
Agric. & fishery	586	0.01	7.08	3.0	218.15	210.36	254.02	0.25	38.83

Notes: Rows are sorted by median hourly earnings, including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

plausible value variables provided in the dataset for each score. Table 2 and Figure 3 show that the three distributions are roughly similar and follow a Gaussian curve with a slight negative skewness. However, it is worth noting that problem-solving data is available only for 71% of the respondents.

Table 3 presents frequencies for the occupation groups and occupation-specific averages for earnings, obtained qualification and skill variables. The occupation groups are mainly based on the 1-digit ISCO codes, with the managers group split into the high and low-skilled ones using the 2-digit codes. This is due to the fact that the two subgroups have different ISCO skill level requirements. The armed forces occupations are omitted due to a small occurrence in the data. Furthermore, since some labour mismatch measures are distribution-based, the respondents that belong to country-specific groups with less than 30 observations are removed from the analysis. The number of observations belonging to each of the occupation groups varies from 554 to 14,882. This could potentially be improved by the wider use of 2-digit and 3-digit codes. However, it may lead to a substantial loss in the number of observations because lower occupation subgroups contain more missing values. The table shows that occupation groups with lower median earnings tend to have lower median qualifications and mean skill scores.

Table 4 has the same structure as Table 3 but presents the statistics across countries. The country-specific sample size has a smaller variance than the occupation-specific one. However, such countries as Canada, Peru, Hungary, Singapore, Germany, Turkey, Austria, the United States, and Sweden were dropped due to the earnings data being available only in deciles. In addition, Italy, France and Spain lack data on problem-solving, and Estonia reports no data on the highest obtained qualification. The table suggests that, unlike occupation, country-specific averages for qualification and skill scores do not feature an obvious association with median earnings.

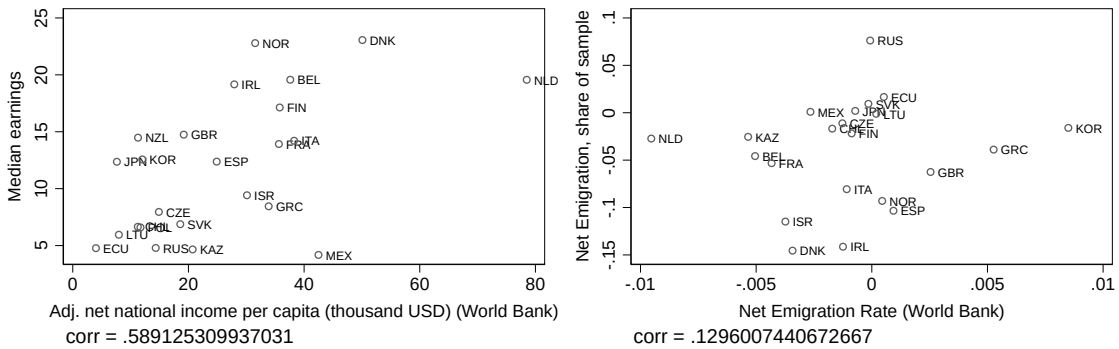
Some derived variables require the calculation of a statistic within certain groups

Table 4: Frequencies and averages: countries

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Denmark	4426	0.06	23.07	7.0	274.42	283.81	283.13	0.51	43.29
Norway	3114	0.04	22.78	8.0	285.89	287.77	291.09	0.51	39.52
Netherlands	3125	0.04	19.63	6.0	292.23	288.20	291.16	0.51	39.90
Belgium	2688	0.04	19.53	7.0	281.33	286.04	283.07	0.50	40.17
Ireland	2724	0.04	19.17	8.0	276.65	266.96	281	0.57	39.15
Finland	3115	0.04	17.13	10.0	298.13	292.67	293.64	0.52	41.30
United Kingdom	4680	0.06	14.78	6.0	280.46	271.16	283.47	0.59	39.73
New Zealand	3521	0.05	14.52	8.0	283.93	273.46	289.44	0.56	NaN
Italy	1762	0.02	14.16	6.0	259.81	260.60	NaN	0.49	41.46
France	3572	0.05	13.92	6.0	268.32	263.42	NaN	0.50	41.08
Korea	3000	0.04	12.64	6.0	274.85	267.31	285.04	0.46	39.68
Japan	3186	0.04	12.44	11.0	300.68	293.02	297.96	0.49	41.52
Spain	2430	0.03	12.38	6.0	259.57	255.52	NaN	0.49	39.82
Israel	2507	0.03	9.49	6.0	257.17	254.49	271.41	0.51	37.11
Greece	1174	0.02	8.48	6.0	257.60	259.97	259.28	0.53	39.24
Slovenia	2168	0.03	8.31	6.0	261.26	264.55	266.51	0.50	41.24
Czech Republic	2566	0.04	8	6.0	279.93	281.20	285.80	0.52	38.63
Estonia	3662	0.05	7.36	NaN	277.64	274.34	274.46	0.58	41.03
Slovak Republic	2419	0.03	6.88	6.0	279.28	284.11	280.73	0.51	40.48
Poland	3851	0.05	6.67	6.0	276.55	268.23	278.20	0.44	31.08
Chile	2309	0.03	6.63	6.0	225.65	211.16	252.83	0.50	38.65
Lithuania	2680	0.04	6.01	9.0	271.36	273.80	260.40	0.61	42.58
Ecuador	1654	0.02	4.80	6.0	195.59	188.42	227.87	0.41	35.78
Russian Federation	1468	0.02	4.79	12.0	285	279.78	287.85	0.63	36.31
Kazakhstan	2534	0.03	4.66	9.0	255.27	251.43	266.32	0.57	38.50
Mexico	2228	0.03	4.22	3.0	226.17	216.68	257.25	0.39	35.87

Notes: Rows are sorted by median hourly earnings, including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

Figure 4: Earnings and Migration: PIAAC and World Bank data compared



of respondents (e.g. RM is computed using the occupation-specific mean and standard deviation of education). For such variables, a minimum group size of 30 is applied to avoid outliers. Similarly, for the variables that only have variation across groups of respondents, such as net migration, observations are assigned missing values if a group contains less than 30 respondents. Furthermore, the dataset is complemented with the World Bank (2023) data for migration. Figure 4 suggests that although the two datasets contain similar country-level data for earnings, their data for migration is considerably different. Thus, Word Bank's migration data is used for the country-level analysis, whereas PIAAC data is used for the analysis at the individual level.

6 Measurement frameworks output and selection

Section 3 reviews a variety of skill and educational mismatch measurement frameworks. Although the list features only six major measures (job analysis, realised matches, indirect and direct self-assessments, Pellizzari-Fichen, Allen-Levels-Van-der-Velden), some frameworks require the researcher to choose values for the parameters determining the resulting mismatch classification (e.g. measure of central tendency and number of standard deviations for RM). It is impractical for the purposes of this study to include the full variety of specifications in the econometric analysis. Therefore, a preliminary investigation is required teal to select the relevant measures for the analysis of the Mincer function. This is done in three steps. Firstly, country-specific labour shares of under, well and over-matched workers are mapped on a colour spectrum, where the position of a data point is determined by its place in the distribution, and the countries are ranged by their median earnings. Secondly, the correlation matrices are presented in a similar fashion. Finally, log-earnings are regressed on a list of controls and all specifications of the mismatch measures, using the least absolute shrinkage and selection operator (Lasso). The graphical analysis aims to reveal country-level patterns between earnings and mismatch levels and identify the degree of similarity among the classification results, whereas the Lasso is used to choose the specifications that are most useful for predicting earnings by selecting the model that minimises the Mean-Squared Prediction Error (MSPE) and eliminate the rest. The final part of this section reviews the heterogeneity in the output of the selected mismatch measures across multiple worker characteristics to provide directions for further research.

6.1 Mismatch shares cross-country comparison

Let us begin the discussion with a brief overview of the mismatch measures that are selected for further analysis as the result of this section. Figure 5 presents country-specific shares of under, well and over-matched workers that were computed using three

educational mismatch measures (JA, RM, and ISA) and three skill mismatch measures (PF-literacy, numeracy and problem-solving), where the countries are sorted by their median earnings in descending order. The key takeaway of this figure is that educational and skill mismatch may exhibit an opposite relationship with earnings. Specifically, the shares of well-matched, according to JA, are generally larger for the countries with lower median earnings. A similar pattern can be spotted in the shares of well-matched computed with RM. ISA does not feature the pattern due to the countries in the top quartile of the earnings distribution having high well-matched shares, but the countries with relatively low well-matched shares are still clustered in the 3rd quartile. On the other hand, the PF measure applied to each of the three PIAAC skill measures produces shares that decrease as the countries' median earnings decrease. A similar trend can be spotted in the shares produced by ALV in the lower quarter of the median earnings distribution. These patterns are reflected in the shares of under and over-matched workers to various extents.[13] Although the statistical significance of these associations is questionable, the fact that educational and skill mismatch is often described in the literature as equivalent makes it sufficient to start the investigation.

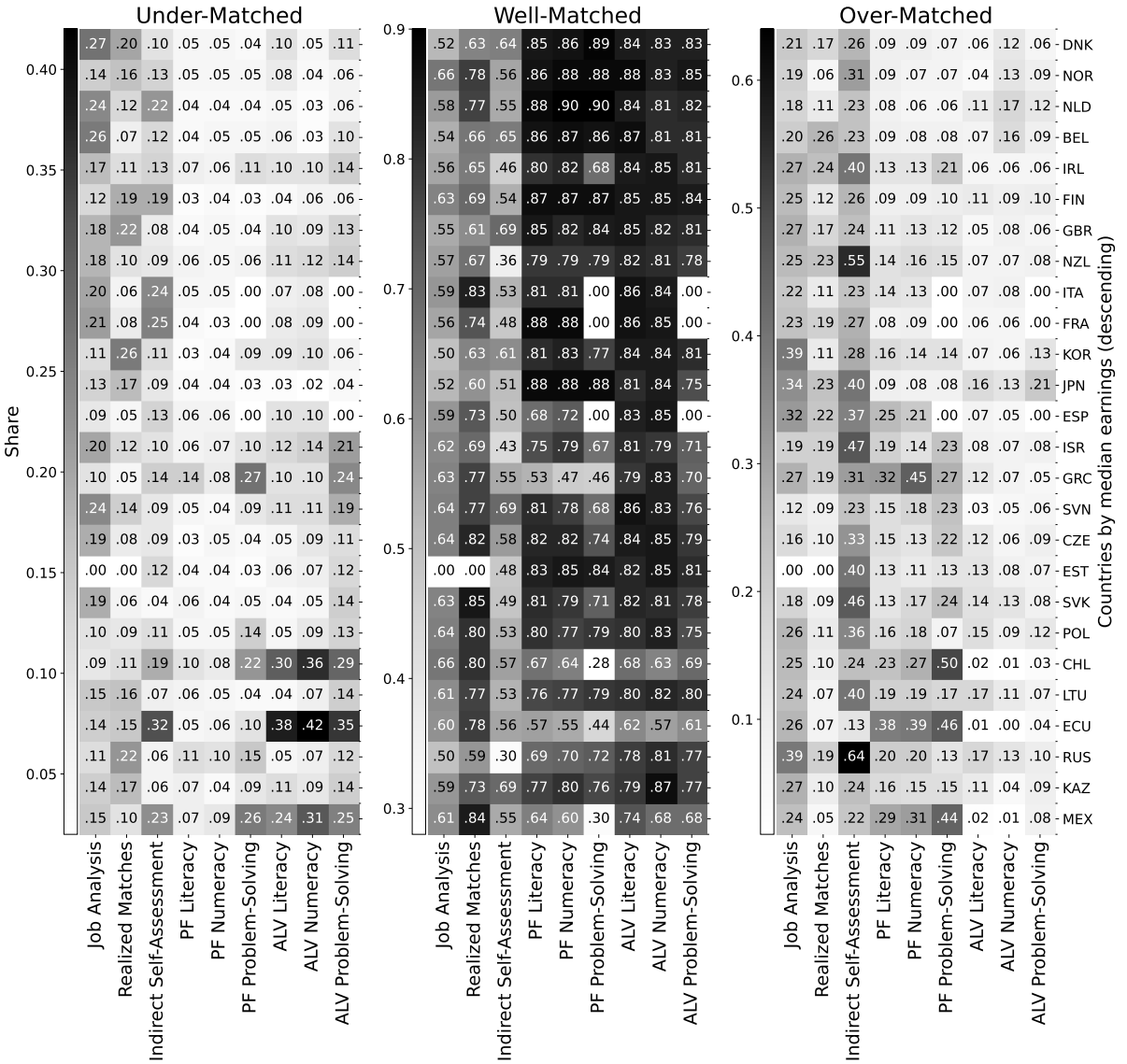
The rest of the section reviews each of the six measures above and their alternative parameter specifications. Notice that JA, which compares workers' ISCO skill levels derived from their highest qualification and ISCO skill level requirement for corresponding occupation group defined by the International Labour Office (2012), is the only measure that does not have alternative specifications. The rest of the measures feature alternative parameter settings that require careful consideration.

Figure 7 maps the shares produced by RM using the information on workers' highest obtained qualifications (ISCED) converted into ISCO skill levels and current job occupation groups.[14] In Section 3, we saw that the literature features multiple specifications of RM, which differ in the measure of central tendency and the number of standard deviations used to determine the classification thresholds. The figure shows that for the mean-based RM, a switch from the most commonly used one standard deviation cutoff to either 0.5 or 1.5 standard deviation threshold results in a considerable difference between the shares, e.g. UK over-education varies between 5% and 29%. In contrast, the mode-based RM is less sensitive to the threshold value despite the coefficient on standard deviation varying across a wider range of 0.1 and 2 as opposed to 0.5 and 1.5 in the case of mean-based RM. The corresponding values for the UK vary between 10% and 22%. This is due to the mode's lower sensitivity to outliers, as well as the nature of the ISCO skill levels metric, which has only four categories. Therefore, the effect of change in the threshold value on the mode-derived mismatch shares is less smooth. The resulting shares

[13] Among other countries, Chile, Ecuador, and Mexico could be of specific interest to the researchers due to their relatively high levels of ALV-computed under-skilling and PF-computed over-skilling.

[14] See Appendix B.

Figure 5: Selected measures: country-specific shares



Notes: Rows are sorted by the country-specific median hourly hourly earnings, including bonuses (PPP corrected USD).
Selected parameters:

RM (Realised Matches): mode ±1 SD cutoff
ISA (Indirect Self-Assessment): 1 year gap
PF (Pellizzari-Fichen): 5th and 95th percentile critical values
ALV (Allen-Levels-Van-der-Velden): 1.5 points difference in the z-scores

for the mean-based specifications of RM show no clear association with median earnings. However, for the mode-based ones, the majority of countries in the upper half of median earnings distribution (Denmark, Belgium, Ireland etc.) feature relatively low mode-based shares of well-educated workers, which is mainly reflected in the over-education and, to a lesser extent, in under-education shares.

Figure 8 presents mismatch shares computed with ISA.[15] This measure compares the number of years of formal education required to get a worker's current job to their highest level of education obtained imputed into years of education. The resulting mismatch classification is determined by the gap between the years of obtained and required education that allows a worker to be labelled as "well-educated". The results are computed for the gaps of 1 to 5 years. As expected, the shares of well-matched are higher when a wider gap is used, e.g. 69% are classified as well-educated in the UK with a gap of 4 years, and 91% with a gap of 1 year. Although the clustering of countries with lower mismatch shares in the 3rd quartile of the median earnings distribution is not obvious in Figure 5, comparing the shares across alternative gap sizes makes it more visible.

The mismatch shares in Figure 9 are computed using DSA. This measure classifies workers based on their answers to two questions: (i) "Do you feel that you have the skills to cope with more demanding duties than those you are required to perform in your current job" and (ii) "Do you feel that you need further training in order to cope well with your present duties". A respondent is considered over-skilled if they answer positively to (i) and negatively to (ii), under-skilled – negatively to (i) and positively to (ii), and well-skilled if they answer negatively to both questions. Column "DP Error" (double-positive error) in Figure 9 contains the share of the respondents who answered positively to both questions. A safe way to interpret these observations is to consider them a measurement error and exclude them from the analysis. However, as shown in the figure, the shares of such responses reach over a quarter of the sample for the majority of the countries. Therefore, in certain cases, the researchers may be tempted to classify the DP respondents as well-skilled, arguing that these respondents are likely to be unsure of their skill potential and, hence, are either well-matched or borderline cases. Thus, it is worth considering this interpretation as a separate skill mismatch measure, which in this paper is referred to as "relaxed" DSA. As implied above, it requires the following assumption

Assumption 1 (Relaxed DSA Homogeneity)**.** *The pooled distribution of the respondents with positive answers to both (i) and (ii) and the respondents with negative answers to both (i) and (ii) is homogeneous.*

One reason why relaxed DSA could potentially be both interesting and problematic in the case of PIAAC data is the potential association between the DP errors and the

[15] See Appendix B.

country-specific median earnings. Figure 9 suggests that the relative number of the DP respondents increases as the median earnings increase. It is also worth noticing that, unlike the output of the education-based measures, the shares of well-matched computed with DSA display a positive relationship with median earnings. However, once the DP observations are merged with the well-matched in relaxed DSA, the large number of DP errors alters the association between the well-matched and earnings, which gains the negative sign. Although this may be considered another result of this section that could be important for the users of PIAAC data, it has limited economic applications because the theoretical reason behind the association is unclear. Furthermore, it hints at the violation of Assumption 1, which would make relaxed DSA unusable for the purposes of this study.

The first 18 columns in Figures 10, 11 and 12 show the shares of well, under and over-skilled workers, respectively, computed with the Pellizzari-Fichen framework. The measure is applied to each of the three sills. Additionally, the shares are computed using both regular and relaxed DSA: regular PF employs the classification thresholds based on the skill distributions of only those respondents who gave negative answers both to (i) and (ii), and relaxed PF utilises the respondents who gave the same answer to both questions (either positive or negative). Finally, the original Pellizzari and Fichen (2017) classification is based on the 5th and 95th percentiles of the well-skilled workers' skill distribution. This specification is referred to as the 10% one, and its output is compared to the 5% and 20% versions of the PF framework that are based on the {2.5th, 75.5th} and {10th, 90th} percentiles, respectively.

Let us first consider the output of regular PF. The effect of the alternative classification thresholds is similar to the one in the cases of ISA and RM: "stricter" specifications produce higher under and over-matched shares and consequently lower well-matched shares. However, unlike ISA and RM, the PF measure yields mismatch shares that have a negative association with the country-specific median earnings. Figure 10 shows that the countries with higher median earnings tend to have a larger share of well-skilled workers. The results are generally consistent across the three skills except in a few countries for which the problem-solving well-matched shares are considerably different from the literacy and numeracy ones, e.g. Chile and Mexico. As shown in figures 11 and 12, this pattern is mainly supported by the results for over-skilling and, to a lesser extent under-skilling: the share of over-skilled workers is negatively associated with the country-specific median earnings. It is worth noting, however, that this pattern is not exhibited by the shares produced with the relaxed versions of PF. On the contrary, relaxed shares display noticeably low variance across counties compared to the output of all other labour mismatch measures. This is due to the fact that the DP errors appear to correlate negatively with the median earnings (see Figure 9). Recall the difference between the regular and relaxed

PF. The pool of well-skilled workers, whose skill distribution is used to compute the classification thresholds, is widened by the respondents with double positive answers. This appears to increase the variance in the skill distributions, which leads to wider thresholds and, hence, lower under and over-skilled shares and higher well-skilled shares. Since the countries with the lower median earnings have a share of errors, it affects their PF shares to a greater extent. The reason behind the negative correlation between the share of the DP respondents and median earnings, however, stays unclear.

The last set of specifications presented in Figures 10-12 is produced by ALV. Similarly to the previous measures, the specifications allowing for larger differences in the z-scores classify fewer workers as mismatched. The shares of the well-matched are mostly similar across countries except for the lowest quarter of the median earnings distributions, where the shares exhibit a similar tendency to PF. In the under-skilling shares, this pattern is less noticeable, with the higher values concentrating around the middle of the median earnings distribution. It is worth mentioning that three countries (Chile, Ecuador, and Mexico) exhibit particularly high shares of the under-skilled. Interestingly, these countries also have the highest shares of DP errors (see Figure 9). Since DSA and AVL are not based on any common variables, this may potentially indicate a wider measurement error issue. Consequently, the shares of the over-skilled are relatively small for the three outlier countries. The rest of the countries reflect a pattern opposite to the one of the under-skilled shares with the higher values concentrated in the tails.

6.2 Correlation analysis

Finally, Figure 13 contains correlation matrices for a subset of the labour mismatch measures considered above. Namely, for RM, the mode ± 1 SD classification threshold is selected due to its robustness to outliers and wide use in the literature. Similarly, the 1-year gap for ISA, 5th and 95th percentiles for the PF critical values, and a 1.5-point z-score difference allowance for ALV are adopted due to its common use among the researchers. As noted above, alternative thresholds mainly affect the "strictness" of the mismatch measure and representativeness of the output mismatch shares. However, it is still worth comparing the performance of regular and relaxed measures due to the association of the DP errors with the country-specific median earnings. The figure suggests that there are three main clusters in the correlation matrix: education-based measures, skill-based measures, and DSAs. The results of JA's classification are the most compatible RM, specifically in the case of under-education, where the correlation reaches 70%. JA also has a relatively high correlation with ISA, reaching 25% for over-matching, but the rest of the measures provide more conflicting results. RM has a similar level of association with ISA, achieving 21% for over-matching and a weaker level of association with the rest of the measures. ISA only appears to have a mild to no association with both skill-

based measures and DSA. In the skill-based sector, each of the three regular PF measures exhibits the highest level of correlation with its relaxed counterpart: 70%, 74% and 62% correlation between the under-matched classifications of literacy, numeracy and problem-solving PF measures, respectively. Both PF mismatch in literacy and numeracy display a considerably stronger association with each other than with the one in problem-solving: 58% against 41% and 38%, respectively, for the well-matched. However, ALV exhibits a considerably lower correlation both with its PF counterparts (22%, 16%, 19% correlation between literacy, numeracy, problem-solving ALV and the respective PF specifications) and among its own specifications applied to different skill variables (29%, 22%, 11% between literacy and numeracy, literacy and problem-solving, numeracy and problem-solving AVL specifications). Lastly, both versions of DSA classifications demonstrate low correlation levels, with all presented alternatives varying between -4% and 6%. This is not surprising given that DSA often classifies the majority of the respondents as over-matched, whereas the rest of the measures label most respondents as well-matched.

Table 5: Frequencies and averages: DSA

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
DP Error	20949	0.29	10.82	8.0	268.50	264.49	277.56	0.49	37.88
Over	39985	0.55	12.26	6.0	273.55	270.57	280.60	0.51	39.40
Under	4891	0.07	13.45	8.0	280.11	275	285.51	0.57	40.08
Well	6738	0.09	13.52	6.0	266.51	262.84	275.38	0.57	44.35

Notes: Earnings – hourly earnings including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

The fact that the output classifications of both regular and relaxed DSA do not correlate with the output of any other measure makes it hard to attribute DSA to either the educational or skill mismatch measures. This creates a challenge for drawing inferences from the comparison between the results of econometric analysis computed using DSA and the results computed with other measures. Furthermore, a large number of DP errors makes it hard to argue that regular DSA uses the same sample as the rest of the measures. This could be addressed using relaxed DSA, where DP errors are labelled as well-matched, assuming the relaxed DSA homogeneity is satisfied. However, Table 9 suggests that Assumption 1 could be challenging to justify. Well-skill respondents exhibit median earnings of \$13.52/hour, whereas the DP respondents' median earnings of \$10.82/hour are below the level of over and under-skilled ones. Furthermore, the two groups appear to have the same difference in the median qualification as the over and under-skilled workers (ISCED level 6 vs level 8, respectively). Finally, the workers that are well-matched in DSA constitute the oldest of the four groups with a mean age of 44.35, whereas the workers with double positive answers – the youngest with a mean age of 37.88. Therefore, it is difficult to argue that the DP respondents are similar to

the well-matched ones. This suggests that even though regular and relaxed DSA could serve as valuable predictors from the statistical point of view, these measures would allow drawing no inference on the effect of labour mismatch.

Although DSA's classification of over and under-matched may be deemed unreliable, its definition of well-matched workers may still be used as input for another measure. The regular and relaxed PF measures produced alternative classifications but correlated enough to form a class of skill mismatch measures. However, the economic interpretation of the difference between the results of the regular and relaxed measures relies on our understanding of the DP respondents, which requires a separate investigation. Therefore, the use of relaxed PF measures in further analysis is postponed. This brings us to the final selection of the six labour mismatch measures featured at the beginning of the section.

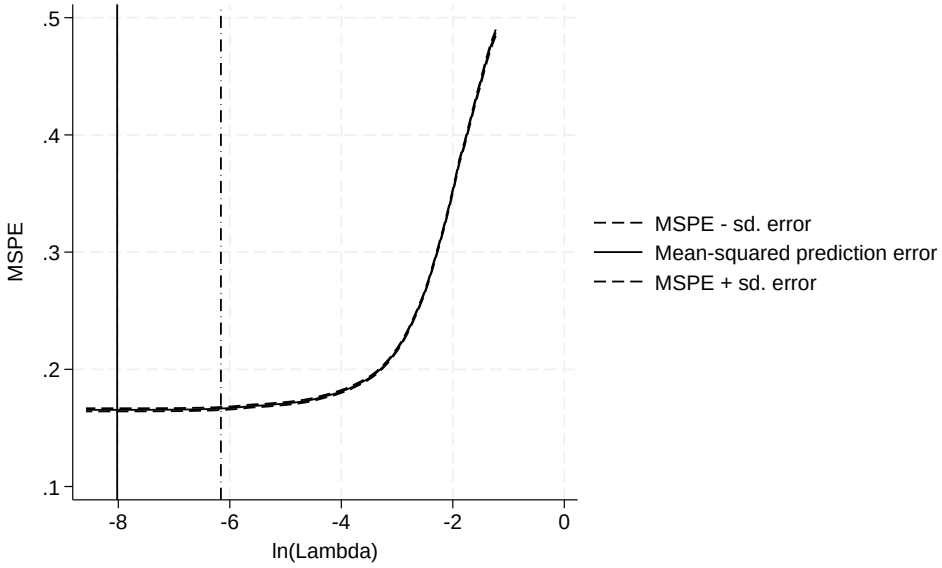
6.3 Out-of-sample prediction performance

This section compares the measures of labour mismatch by their out-of-sample (OOS) prediction performance. This is done by feeding the considered measure specifications to Lasso (Frank and Friedman, 1993; Tibshirani, 1996), a regularised regression model. Using the formulation outlined in Ahrens et al. (2020), the Lasso estimator is derived by minimising the mean squared error subject to the overall and predictor-specific penalties on the absolute value of the coefficient estimates (l_1 regularisation), denoted by λ and ψ, respectively:

$$\hat{\boldsymbol{\beta}}_{lasso}(\lambda) = \arg\min \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \frac{\lambda}{n} \sum_{j=1}^{p} \psi_j |\beta_j|. \tag{6}$$

This often leads the Lasso to set some of the p estimates to zero, thereby making it a model-selection tool. Since the set of selected covariates is determined by λ, it is tuned to optimise the OOS prediction performance using the cross-validation (CV) procedure (Geisser, 1975) with 10 folds. Specifically, each of the 10 folds that the data is split into, in turn, serves as the validation set. The model is estimated using 9 other folds, after which the mean squared prediction error (MSPE) is computed by comparing its prediction for the validation set and the true values. The value of λ that yields the smallest MSPE, λ_{lopt}, is then selected for the analysis (see Figure 6). Alternatively, a more parsimonious model can be selected by utilising the largest λ within one standard deviation of the one that minimises the MSPE, λ_{lse}. Furthermore, since multiple mismatch measures exhibit a high degree of correlation between each other, the model is estimated using the adaptive lasso (Zou, 2006), which utilised the penalty loadings of $\psi_j = 1/|\hat{\beta}_{0,j}|^\theta$ and requires less restrictive assumptions regarding the correlation between the predictors.

Figure 6: Lasso's Mean-Squared Prediction Error over $\ln(\lambda)$



Notes: The solid and dashed vertical lines correspond to λ_{lopt} and λ_{lse}, respectively.

The Lasso is applied to the following regression equation[16]:

$$\ln w_i = \alpha + \boldsymbol{\rho} \times \mathbf{MismatchMeasures}_i + \boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i, \tag{7}$$

where $\ln w_i$ is the natural log of earnings, the first set of covariates in $\mathbf{X}_i$ contains the controls that are commonly included in the Mincer equation; the second set contains variables controlling for migration, the third set accounts for countries and industries fixed effects by incorporating the corresponding sets of dummies in the model, the rest of the covariates refer to the variables that are used for computing educational and skill mismatch, respectively. Finally, $\mathbf{MismatchMeasures}_i$ contains all mismatch measures and their specifications considered in the analysis. It is worth noting that the base category for each mismatch measure is well-matched. Thus, $\boldsymbol{\rho}$ corresponds to the vector of coefficients for the under and over-matched. The purpose of this exercise is to identify the non-zero coefficient estimates in $\hat{\boldsymbol{\rho}}$, i.e. the mismatch measures that are selected by the Lasso to achieve optimal OOS prediction performance.

[16]The vector of covariates $\mathbf{X}_i$ includes personal characteristics ($Female$, Age, Age^2, $Tenure$), migration controls ($MigratedAfter16$, $YearsInCountry$), country and industry fixed effects ($CountryCode$, $IndustryCode$), variables used in the measures of educational mismatch ($Education$, $IscoRequired$, $YearsAtSchool$, $YearsToGetJob$), variables used in the measures of skill mismatch ($NotChallenged$, $NeedTraining$, $Literacy$, $Numeracy$, $ProblemSolving$, $LiteracyUse$, $NumeracyUse$, $ProblemSolvingUse$).

The vector of the regressors of interest $\mathbf{MismatchMeasures}_i$ includes job analysis, realised matches (mean-based with 0.5, 1 or 1.5 SDs thresholds or mode-based with 0.1, 1 or 2 SDs thresholds), direct self-assessment (regular or relaxed), indirect self-assessment (1-5 year gaps), Pellizzari-Fichen (regular or relaxed, literacy, numeracy or problem-solving based; 0.025, 0.05 or 0.1 quantile thresholds), Allen-Levels-van-der-Velden (literacy, numeracy or problem-solving based; 1, 1.5 or 2 z-score gaps).

Table 16 presents the coefficient estimates for the three Lasso specifications.[17] Adaptive Lasso selects the most encompassing model. In addition to the majority of the controls, it incorporates JA, both DSA specifications, ISA with different gaps, a variety of mean and mode-based RM, as well as numerous PF and ALV specifications. Overall, 142 out of 144 are selected. The regular Lasso with λ_{lopt} results in a more conservative model. Although most of the controls are still selected, the DSA is completely excluded, and most of the other measures lose specifications that are only different by the strictness of the classification thresholds. Finally, the most parsimonious model is selected by the Lasso utilising λ_{lse}. This excludes some controls. Only leaving one or two specifications per measure per mismatch type (i.e. under and over-matched). The Lasso appears to give slight preference to the stricter versions of ISA, mode-based RM over the mean-based, regular over relaxed PF and only allows a few non-zero coefficients for the AVL. Overall, the results suggest a valuable contribution of the labour mismatch measure to the OOS prediction performance of the Mincer earnings function. Despite including all the measure components among the controls, even the least generous Lasso does not fully exclude any of the measures except DSA. This suggests that there are no strong data-driven reasons to alter the selection made in this section so far.

6.4 Heterogeneity

The final part of this section presents the graphical heterogeneity analysis of earnings and labour mismatch across the main workers' characteristics: gender, age, migration, education, and skills.[18] The analysis is conducted by comparing the distributions of mismatch shares for the workers that share specific characteristics. Since the country-specific share would not provide a sufficient number of groups, the shares are computed for 346 markets, where a market is defined as a country-specific industry with a minimum of 30 workers. The mismatch shares are calculated using the set of selected measures: JA, mode-based RM with 1 SD thresholds, 1-year gap ISA, as well as 0.05 percentile PF and 1 z-score AVL applied to literacy, numeracy, and problem-solving scores.

To begin with, Figures 14 and 15 demonstrate the distributions of the log-earnings for various categories of workers. The top graph of Figure 14 suggests a negative wage gap between male and female workers. The middle plot illustrates the positive wage gaps between the older (over 45 y.o.) and the middle-aged (30 to 44 y.o.) workers, as well as between the middle-aged and younger (under 30 y.o.) workers. The last plot in the figure suggests a positive wage gap between the migrated and local workers. Figure 15 presents a similar breakdown by the workers' education (measured using the four ISCO skill levels) and PIAAC skill scores categorised into quartiles. The top graph shows

[17] See Appendix D.1.
[18] See Appendix C.

positive wage gaps between each of the two consequent ISCO Sill Levels, suggesting that more educated workers tend to earn higher wages. A similar picture is observed in the other three plots corresponding to literacy, numeracy and problem-solving. These results are only inconsistent for the workers who are placed in the 4th quartile of the problem-solving distribution and do not appear to have higher earnings than those placed in the 3rd quartile.

Figures 17 and 18 contain the shares of under and over-matched workers by gender. The top three graphs in Figure 17 exhibit negative gender mismatch gaps, suggesting that women are less likely to be under-educated than men. Similar results are observed in the under-skilling shares produced by ALV. However, the PF plots lead to mixed conclusions with numeracy and problem-solving-based measures suggesting slight positive gender mismatch gaps. The over-matching distributions presented in Figure 18 show that women are more likely to be overeducated. As mentioned above, the results of ALV shares are similar to those of the education-based measures, although they exhibit lower magnitude. Nevertheless, the PF shares demonstrate clear negative gender mismatch gaps, suggesting that women are less likely to be over-skilled. These results show that gender is an important determinant of labour mismatch and may exhibit different associations depending on the mismatch measure of choice.

Another commonly used control that appears to have an association with the mismatch outcomes is age. The top three graphs in Figure 20 show that older workers are more likely to be under-educated than both middle-aged and younger ones. The same pattern is displayed by the under-skilling shares computed using PF and ALV, with the exception of numeracy-based ALV, where the difference in the distributions is less clear. Furthermore, these results are reflected in the shares of over-matched workers presented in Figure 21. The older workers are less likely to be over-educated and over-skilled, which is supported by all measures except RM. This result could be explained by the young workers' struggle to find a job matching their level of skill and education due to the lack of experience. This illustrates the difference between the notions of skill and experience, showing the importance of controlling for both.

Let us now consider the differences in mismatch distributions based on migration. The under-education shares presented in Figure 23 show a negative migration mismatch gap, suggesting that migrated workers are less likely to be under-educated. However, both sets of skill-based measures exhibit positive mismatch gaps. This conflicting pattern is mirrored by the over-matching shares in Figure 24, showing that migrated workers are more likely to be over-educated but less likely to be over-skilled. Similarly, this conclusion is supported by all measures. The higher level of over-education among the migrated workers could be attributed to the language and cultural barriers preventing them from being employed in the jobs that would allow them to utilise their qualifications, as well as

some countries' policies protecting the local workers. Since the respondents are classified as migrated only if they entered a country after turning 16 years old, the higher level of under-skilling could be explained by the differences in the school education systems between the country of birth and the country of residence. This, however, does not explain such a consistent split in the results between the education and skill-based measures.

Finally, Figures 25 to 30 show a limited variation of educational mismatch across the quartiles of the skills' distributions and skill mismatch across ISCO skill levels. This supports the notion that skill and education are separate characteristics, which, although they may correlate on their own, are not necessarily good predictors of each other when put in the context of labour mismatch. Nevertheless, as mentioned at the beginning of this section, both skill and education variables appear to affect earnings. Hence, they should be used as controls in the Mincer equation as long as multicollinearity is avoided.

7 The market-level error components model

The statistical analysis is based on Verdugo and Verdugo's (1989) version of the ORU Mincer earnings function as defined in equation (3). To address the concern regarding unobserved heterogeneity raised in Section 6, the equation is modified using an error components model. Suppose $N = \sum_{j=1}^{M} n_j$ workers i are employed in M labour markets j, where j is defined as a country-specific industry as per the International Standard Industrial Classification of All Economic Activities Revision 4 (ISIC). Equation (3) can then be written as follows.

$$\ln w_{ij} = a + \rho_o over_{ij} + \rho_u under_{ij} + \boldsymbol{\beta}\mathbf{x}_{ij} + \mu_j + \eta_{ij}, \tag{8}$$

where $\ln w_{ij}$ is the natural logarithm of earnings, $over_{ij}$ is a binary variable that takes the value of 1 if worker i is over-matched and 0 otherwise, $udered_{ij}$ is defined in a similar fashion, $\mathbf{x}_{ij}$ is a vector of controls, μ_j is market-level unobserved heterogeneity, and η_{ij} is idiosyncratic error such that $E[over_{ij}\eta_{ij}] = E[under_{ij}\eta_{ij}] = 0$. Since we suspect that μ_j could potentially be correlated with the regressors of interest, estimating equation (8) with pooled ordinary least squares (POLS) allows the $\hat{\rho}$s to be driven by unobserved heterogeneity, likely leading to biased and inconsistent estimates. Nevertheless, $\hat{\rho}_o^{POLS}$ and $\hat{\rho}_u^{POLS}$ are still valuable as they capture the variation both within and between the markets, hence, constitute the 1st specification (the base model) for the analysis.

To obtain more appropriate estimates, we need to correct for μ_j. Following Mundlak (1978), let us define the relationship between the unobserved heterogeneity and regressors as

$$\mu_j = \phi_o \overline{over}_j + \phi_u \overline{under}_j + \boldsymbol{\gamma}\overline{\mathbf{x}}_j + \nu_j, \tag{9}$$

where $\bar{\cdot}_j$ denotes an average over i. Substituting this for μ_j in equation (8) yields

$$\begin{aligned}\ln w_{ij} = \alpha &+ \rho_o over_{ij} + \rho_u under_{ij} + \boldsymbol{\beta}\mathbf{x}_{ij} \\ &+ \phi_o \overline{over}_j + \phi_u \overline{under}_j + \boldsymbol{\gamma}\overline{\mathbf{x}}_j + \nu_j + \eta_{ij}.\end{aligned} \tag{10}$$

It can be shown that applying Generalised Least Squares (GLS) to estimate the ρs in equation (10) is equivalent to the fixed effects (FE) estimation of equation (8), i.e. $\hat{\rho}^{FE} = \hat{\rho}^{GLS}$ (Mundlak, 1978). These estimates are valuable for two reasons. Firstly, $\hat{\rho}_o^{FE}$ and $\hat{\rho}_u^{FE}$ capture the individual-level association between labour mismatch and earnings. Secondly, the comparison of the FE and POLS estimates provides an idea of the direction and magnitude of the bias caused by the unobserved market-specific factors. Thus, the FE estimation corresponds to the 2nd specification.

To investigate the variation in the unobserved factors further, it is useful to consider the between estimates (BEs) of the ρs. These can be obtained by summing the GLS coefficient estimates for the variable of interest and its average in equation (10), i.e. $\hat{\rho}^{BE} = \hat{\rho}^{GLS} + \hat{\phi}^{GLS}$ (Mundlak, 1978). It's worth noting that the interpretation of the BEs is different from the ones produced by POLS and FE. Since $over_{ij}$ and $under_{ij}$ are binary variables, $\overline{over}_j$ and $\overline{under}_j$ are the shares of the workers, for whom the respective variables take the value of one. Hence, $\hat{\rho}_o^{BE}$ ($\hat{\rho}_u^{BE}$) corresponds to the average change in earnings associated with switching from a market with no over-matched (under-matched) workers to a fully over-matched (under-matched) one, and constant α^{BE} represents the average log-earnings in a fully well-matched market.[19] Although the ρ^{BE} cannot be directly compared with the ρ^{FE} and ρ^{POLS}, they are important for understanding the market-level association between the mismatch and earnings, therefore, complements the FE in the 2nd specification.

Finally, the analysis employs random effects (RE) as an alternative way to combine the within and between sources of variation. The RE estimates are obtained by applying GLS to the quasi-demeaned version of equation (8):

$$\begin{aligned}\ln w_{ij} - \theta\overline{\ln w}_j = \alpha(1-\theta) + \rho_o(over_{ij} - \theta\overline{over}_j) + \rho_u(under_{ij} - \theta\overline{under}_j) \\ + \boldsymbol{\beta}(\mathbf{x}_{ij} - \theta\overline{\mathbf{x}}_j) + \mu_j(1-\theta) + \eta_{ij} - \theta\overline{\eta}_j.\end{aligned} \tag{11}$$

Unlike POLS, which gives equal weights to the fixed effects and between estimates, the RE estimator produces a matrix-weighted average of the $\hat{\rho}^{FE}$s and $\hat{\rho}^{BE}$s with the inverse of their respective variances as weights (Baltagi, 2008). Since choosing RE over POLS may potentially lead to efficiency gains, the $\hat{\rho}^{RE}$s are presented as the 3rd specification.

[19] An arguably more useful approach is to divide the BEs by 10 and interpret them as an increase in earnings associated with the 10% increase in average over/under-matching.

8 Main results

This section presents the results of estimating Verdugo and Verdugo's (1989) version of the ORU Mincer earnings function, which are summarised in tables 17 to table 25.[20] Each of the nine tables corresponds to one of the selected mismatch measures. The three specifications described in Section 7 (POLS, Mundlak FE, and RE) employ the same set of controls: gender, age, age-squared, tenure, migration, years in the country, market-specific net emigration share, country-specific World Bank net emigration rate, as well as natural logarithms of literacy, numeracy and problem-solving for the models using an education-based measure of mismatch, and ISCO skill level for the models using a skill-based measure. Additionally, the Mundlak FE specification includes the market-specific averages of both the regressors of interest and covariates, which allows the calculation of the BE estimates.

The POLS coefficient estimates suggest that being under-matched is associated with a positive difference in earnings for RM and ISA (4.6% and 5.7%, respectively), no significant difference for JA, a negative difference for the literacy, numeracy and problem-solving-based PF and ALV (-12.6%, -10.9%, -24.2%, -8.2%, -13.8% and -19.2%, respectively). The corresponding 95% confidence intervals are estimated to be $[1\%, 9\%]$, $[2\%, 9\%]$, $[-16\%, -9\%]$, $[-15\%, -7\%]$, $[-30\%, -19\%]$, $[-13\%, -4\%]$, $[-18\%, 10\%]$, and $[-24\%, -14\%]$. Being over-matched is associated with an increase in earnings only for ALVN and ALVP (11% with $[7\%, 15\%]$ CI and 3.5% with $[1\%, 6\%]$ CI) and a decrease for the rest of the measures (varying between -5.1% $[-8\%, -2\%]$ and -14.6% $[-17\%, -12\%]$).

When unobserved heterogeneity is removed in the Mundlak FE specification, the under-matching coefficients for most mismatch measures lose magnitude but maintain their signs and statistical significance. However, the estimates for ISA, PFL, and ALVN become indistinguishable from zero, and ALVL's coefficient switches the sign from negative to positive, suggesting a 4.1% $[2\%, 6\%]$ higher earnings among the under-skilled. A similar pattern of magnitude loss is observed for over-matching coefficients of ISA, PFL, PFN, and ALVL, although a slight increase in the magnitude is displayed by JA. Additionally, $\hat{\rho}_o^{ALVN}$ loses the significance, $\hat{\rho}_o^{RM} = -3.4\%$ gains the significance; $\hat{\rho}_o^{PFL} = 5.4\%$, $\hat{\rho}_o^{PFN} = 6.2\%$ and $\hat{\rho}_o^{PFP} = 2.6\%$ change the sign to positive, and $\hat{\rho}_o^{ALVP} = -3\%$ switches the sign to negative. The CIs for the five updated estimates above amount to $[-6\%, -1\%]$, $[3\%, 7\%]$, $[4\%, 8\%]$, $[0\%, 5\%]$, and $[-5\%, -1\%]$, respectively. It is worth noting that the FE estimates appear to be more precise than the ones of POLS, producing narrower confidence intervals. To summarise, POLS produces positive coefficient estimates for under-education and negative estimates for over-education, under-skilling and over-skilling. The results of FE estimation suggest that unobserved heterogeneity increases the estimates'

[20] See Appendix D.2.

variance and drives them away from zero, although there are exceptions when the bias appears to push the coefficients towards zero or cause a change of the sign.

Let us switch the focus to solely market-level variation, computed by taking the sum of the Mundlak FE coefficient and a coefficient for the corresponding market-level mean. The BE coefficient estimates for ISA suggest that a 10% increase in educational mismatch is associated with a change in average earnings of 19.49% for under-education and −4.47% for over-education. The use of JA and RM leads to no evidence of the effect of either under or over-education on earnings at the market level. The skill-based measures produce mostly negative coefficients for both under-skilling ($\hat{\rho}_u^{PFL} = -17.04\%$, $\hat{\rho}_u^{PFN} = -12.69\%$, $\hat{\rho}_u^{PFP} = -10\%$, $\hat{\rho}_u^{ALVL} = -9.85\%$, $\hat{\rho}_u^{ALVP} = -16.22\%$) and over-skilling ($\hat{\rho}_o^{PFL} = -9.85\%$, $\hat{\rho}_o^{PFN} = -8.98\%$, $\hat{\rho}_o^{PFP} = -3.76\%$, $\hat{\rho}_o^{ALVL} = -8.53\%$), except ALVN, which shows no significant estimate for under-skilling and a 17.55% increase for over-skilling. Overall, the coefficient estimates computed with BE demonstrate the difference between the education and skill-based measures in their association with earnings at the market level.

Finally, the RE results exhibit little to no difference from the FE estimates. Since the $\hat{\rho}^{RE}$s are the matrix-weighted averages of the corresponding $\hat{\rho}^{FE}$s and $\hat{\rho}^{BE}$s (Baltagi, 2008), it can be concluded that the bulk of the identifying variation comes from the individual rather than the market level.

Some additional inferences can be drawn from the RE coefficient estimates for the controls. The coefficient for gender exhibits significant negative estimates for all mismatch measures, varying between −13.6% and −16.2%, which is roughly consistent with the EU gender wage gap as estimated by the European Commission (2022). Being a year older is reported to be associated with higher earnings between 5.6% and 5.8%. Age-squared displays significant negative estimates but with the extremely low magnitude of −0.1% or less, suggesting little nonlinearity. Finally, the change in earnings associated with one additional year of tenure is estimated to be $\sim 1\%$. Migrating after the age of 16 is associated with no significant change in earnings when the mismatch is measured using education-based measures. However, in the models with skill-based measures, migrants are reported to have −17.9% to −22.4% lower remuneration. This suggests a difference in the correlations between migration and the two types of mismatch. Spending an extra year in the country following the migration is reported to lead to a positive change in the earnings of 0.2% to 0.5% regardless of the mismatch measure. Additionally, a 1% increase in market-level net migration share is associated with a decrease in earnings of about −6.5%, and the coefficients for the country-specific net emigration rate derived using the World Bank data feature no estimates that are significantly different from zero. The coefficient estimates for the skill variables suggest that a 10% increase in literacy, numeracy and problem-solving is associated with an increase in earnings of roughly 2%,

4% and 2.5%, respectively, in all models. Finally, improving one's education from ISCO skill level 0 to ISCO skill levels 2, 3 and 4 are reported to result in a $\sim 15.2\%$, $\sim 33.5\%$, and $\sim 54.2\%$ increase in earnings, respectively, varying by a few percentage points across the models. Overall, the results show no evidence of major differences in the behaviour of the controls across the models utilising different measures of labour mismatch, except the binary variable for migration, which demonstrated conflicting estimates.

9 Conclusion and research perspectives

This work aims to investigate the implications of using skill-based and education-based measures of labour mismatch on the results of estimating Verdugo and Verdugo's (1989) version of the Mincer earnings function. The analysis focuses on the mismatch-earnings association at the individual and market levels. To explore this, an error components model is applied to the cross-sectional data for 26 countries provided by the OECD Survey of Adult Skills (PIAAC).

The results suggest that the choices of both an input variable and labour mismatch measure are crucial for the coefficient estimates of simple pooled OLS. Specifically, realised matches and literacy-based Allen-Levels-Van-der-Velden tend to produce positive estimates for under-matching coefficients, thereby supporting Verdugo and Verdugo's (1989) findings of positive returns to under-education, whereas using numeracy and problem-solving-based Pellizzari-Fichen and problem-solving-based Allen-Levels-Van-der-Velden lead to negative estimates. Moreover, the results of fixed effects estimation show that the coefficients are biased by market-level unobserved heterogeneity in different directions, with the skill-based measures being dominantly biased downwards and the education-based measures exhibiting mixed directions. Similarly, the market-level analysis of between estimates demonstrates insignificant coefficients in the RM and JA models, mixed for ISA, and mostly negative ones for the PF and ALV measures.

Given the above, one can draw two main conclusions: (i) education and skill mismatch should be distinguished both conceptually and empirically and, if used as a proxy for each other, are unlikely to produce accurate results in the analysis of the Mincer earnings function, and (ii) the coefficient estimates for under and over-matching are sensitive to the choice of a mismatch measure. The former implies that the researchers ought to be careful about the economic inference they draw from empirical analysis if they analyse a mismatch in qualifications, training or other education-like characteristics. The latter suggests that the results of a mismatch analysis are often specific to the measure of choice, and the interpretation would often depend on the underlying features of the workers and jobs that the measure is focused on. Even though the fact that different mismatch measures may produce conflicting results may appear inconvenient, it provides

an opportunity to extract information about a labour market from different angles using a variety of mismatch measures. However, since the specific features of workers and jobs causing the output of mismatch measures to deviate from one another are unclear, the procedure for using these deviations to draw economic inferences is not obvious.

The heterogeneity analysis provides additional results worth mentioning. Namely, the graphical analysis shows that women are more likely to be over-educated, whereas men are more likely to be under-educated but also over-skilled, when over-skilling is measured using Pellizzari-Fichen. Interestingly, the output of the Allen-Levels-Van-der-Velden produces gender gaps that are more similar to the ones of education-based measures than the fellow skill-based PF measure. Furthermore, we find that the older workers are more likely to be under-matched than the younger workers in both education and skill but less likely to be over-matched. Finally, the migrated workers are less likely to be under-educated and over-skilled but more likely to be over-educated and under-skilled, which is supported by both Pellizzari-Fichen and Allen-Levels-Van-der-Velden. Since the graphical analysis only allows us to identify the patterns in the output of the mismatch measures, a more thorough examination is required to establish the influence that gender, age and migration may have on the coefficient estimates of different mismatch measures in the Mincer equation. This, along with the strategies for extracting additional information from the output of mismatch measures, is left for further research.

A Additional summary statistics

Table 6: Frequencies and averages: JA

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	16890	0.23	11.20	11.0	271.17	265.72	282.32	0.56	38.15
Under	11635	0.16	15.65	6.0	274.75	273.75	280.82	0.51	42.17
Well	40363	0.56	12.09	6.0	270.83	267.44	278.63	0.49	39.10
nan	3675	0.05	7.36	NaN	277.60	274.28	274.48	0.58	41.02

Table 7: Frequencies and averages: RM

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	9988	0.14	13.68	12.0	285.30	282.29	288.30	0.57	38.02
Under	9328	0.13	15.36	7.0	273	270.14	281.06	0.53	42.37
Well	49572	0.68	11.67	6.0	268.54	264.83	277.76	0.50	39.08
nan	3675	0.05	7.36	NaN	277.60	274.28	274.48	0.58	41.02

Table 8: Frequencies and averages: ISA

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	23065	0.32	10.59	9.0	275.48	271.30	281.33	0.54	37.94
Under	9065	0.12	13.93	5.0	262.70	259.51	274.49	0.46	42.50
Well	38384	0.53	13.08	7.0	274.55	271.69	280.40	0.52	39.59
nan	2049	0.03	6.61	3.0	221.91	213.39	252.34	0.45	40.35

Notes: Earnings – hourly earnings including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

Table 9: Frequencies and averages: DSA

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
DP Error	20949	0.29	10.82	8.0	268.50	264.49	277.56	0.49	37.88
Over	39985	0.55	12.26	6.0	273.55	270.57	280.60	0.51	39.40
Under	4891	0.07	13.45	8.0	280.11	275	285.51	0.57	40.08
Well	6738	0.09	13.52	6.0	266.51	262.84	275.38	0.57	44.35

Table 10: Frequencies and averages: PF-Literacy

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	10621	0.15	11.06	10.0	312.28	304.86	311.67	0.47	35.56
Under	3896	0.05	10.55	6.0	196.14	197.09	223.60	0.51	43.22
Well	58004	0.80	12.29	6.0	269.57	266.51	276.95	0.53	39.94
nan	42	0	10.10	6.0	262.92	272.97	276.31	0.38	36.31

Table 11: Frequencies and averages: PF-Numeracy

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	10819	0.15	10.94	9.0	303.28	310.26	306.65	0.40	36.26
Under	3754	0.05	10.70	6.0	204.85	188.36	227.95	0.57	42.57
Well	57948	0.80	12.32	6.0	270.37	265.76	276.82	0.53	39.87
nan	42	0	10.10	6.0	262.92	272.97	276.31	0.38	36.31

Table 12: Frequencies and averages: PF-Problem-Solving

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	7843	0.11	11.11	11.0	301.14	300.65	318.43	0.45	32.98
Under	3797	0.05	9.94	6.0	222.45	221.10	211.39	0.52	41.67
Well	39118	0.54	13.99	9.0	280.60	279.76	278.74	0.54	38.50
nan	21805	0.30	10.15	6.0	254.32	244.65	255.32	0.49	42.96

Notes: Earnings – hourly earnings including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

Table 13: Frequencies and averages: ALV-Literacy

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	6254	0.09	8.77	6.0	311.45	301.87	307.56	0.54	34.82
Under	6791	0.09	11.93	6.0	215.43	215.59	239.90	0.46	41.82
Well	59436	0.82	12.44	7.0	274.18	270.92	281.45	0.52	39.70
nan	82	0	11.43	8.5	264.40	258.36	285.57	0.59	40.30

Table 14: Frequencies and averages: ALV-Numeracy

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	6123	0.08	13.97	11.0	315.54	320.15	309.88	0.51	37.47
Under	7289	0.10	9.74	6.0	222.69	210.72	248.58	0.51	38.92
Well	59095	0.81	12.11	6.0	273.43	270.17	279.46	0.52	39.75
nan	56	0	10.64	7.0	260.97	249.36	279.27	0.52	37.96

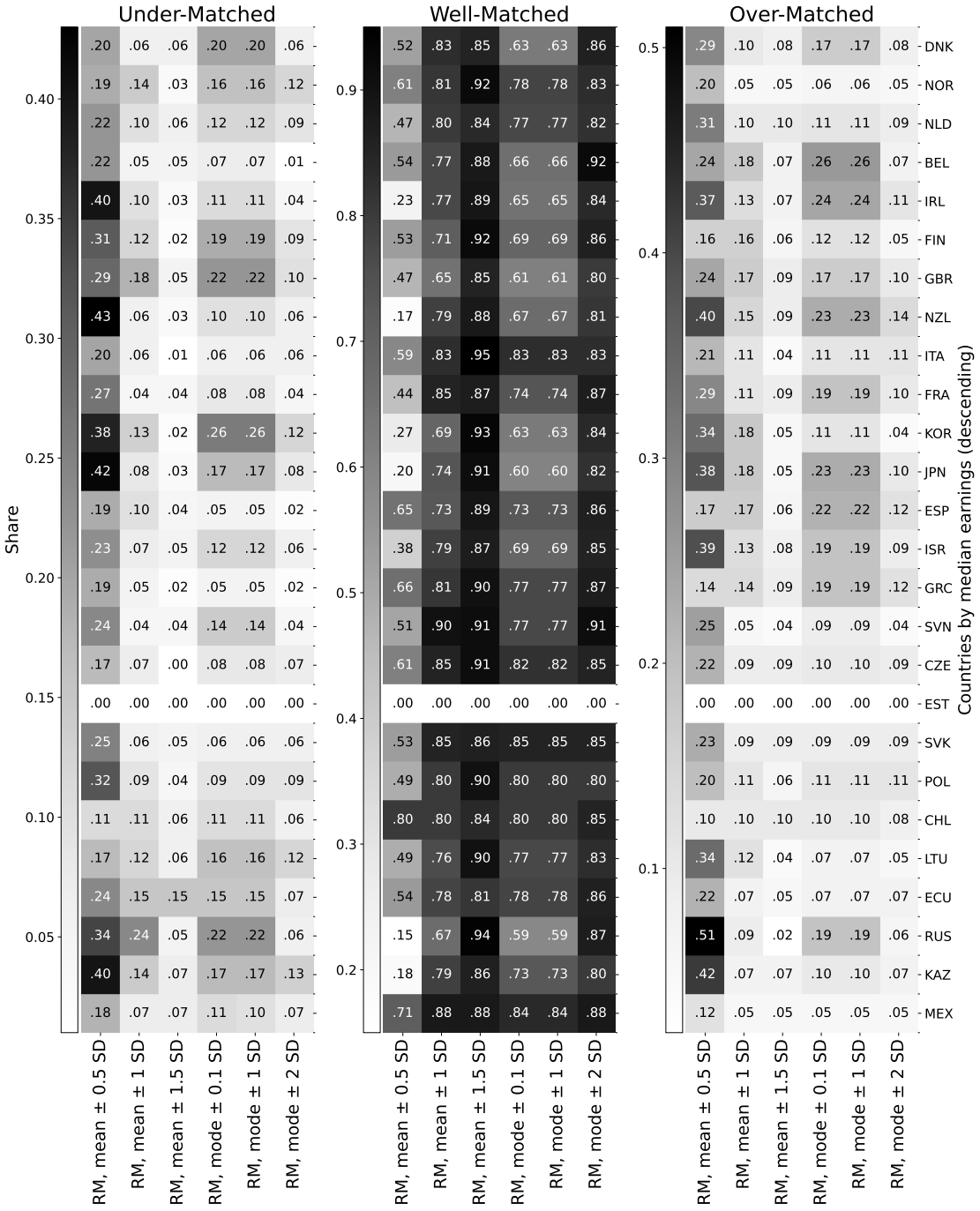
Table 15: Frequencies and averages: ALV-Problem-Solving

	N	Frac.	Median earnings	Median qual.	Mean literacy	Mean num-y	Mean pr. slv.	Mean gender	Mean age
Over	4459	0.06	11.37	7.0	307.91	305.83	320.64	0.55	31.99
Under	6481	0.09	10.97	6.0	235.84	235.31	227.36	0.51	41.59
Well	40216	0.55	13.72	9.0	282.89	282.12	283.52	0.53	37.92
nan	21407	0.30	10.24	6.0	254.60	244.84	256.77	0.49	43.10

Notes: Earnings – hourly earnings including bonuses (PPP corrected USD). Qualification (qual.) – ISCED 1997 level.

B Mismatch measures output

Figure 7: Realised matches: country-specific shares

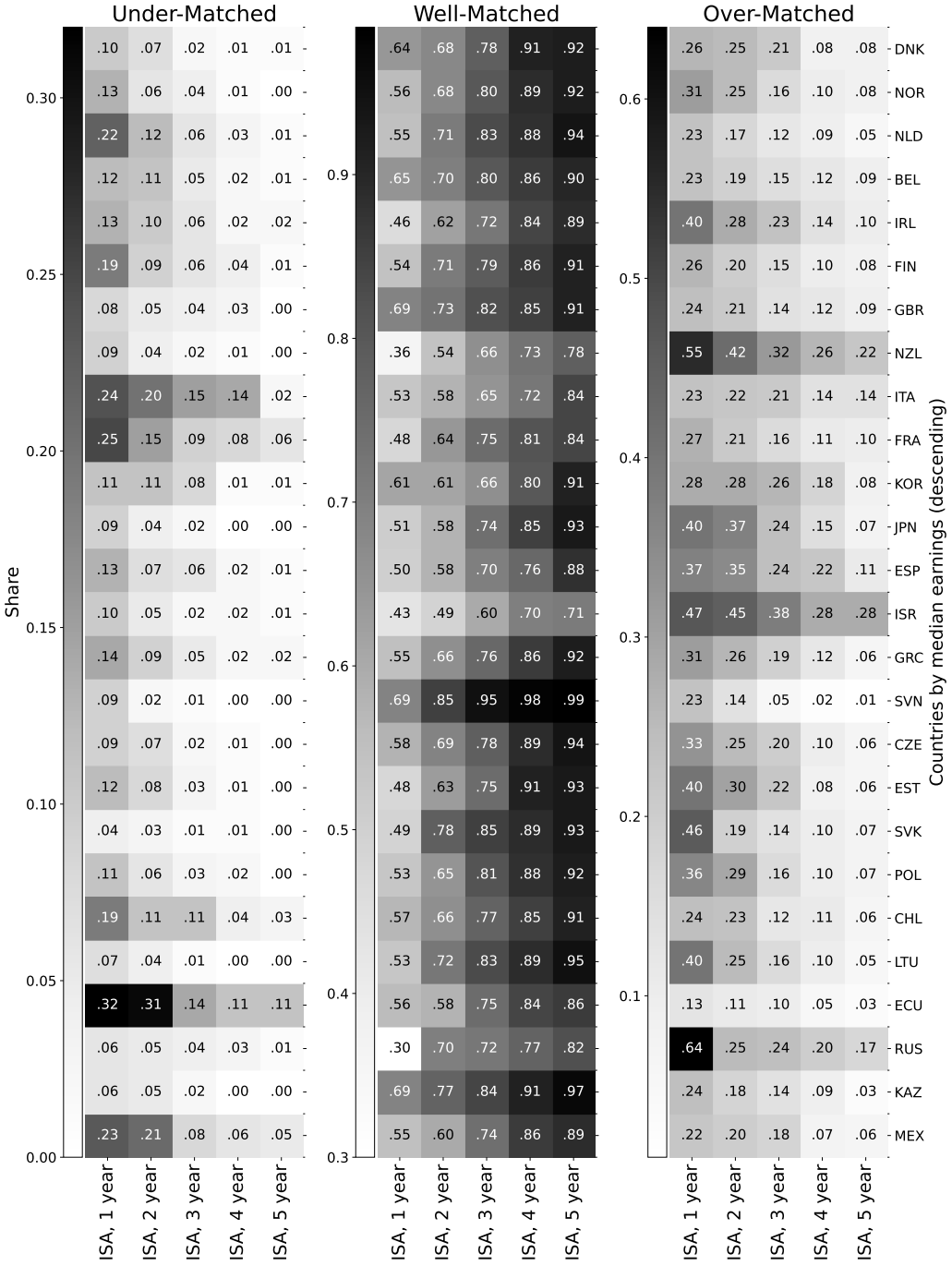


Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD). Education data is missing for Estonia.

C Heterogeneity

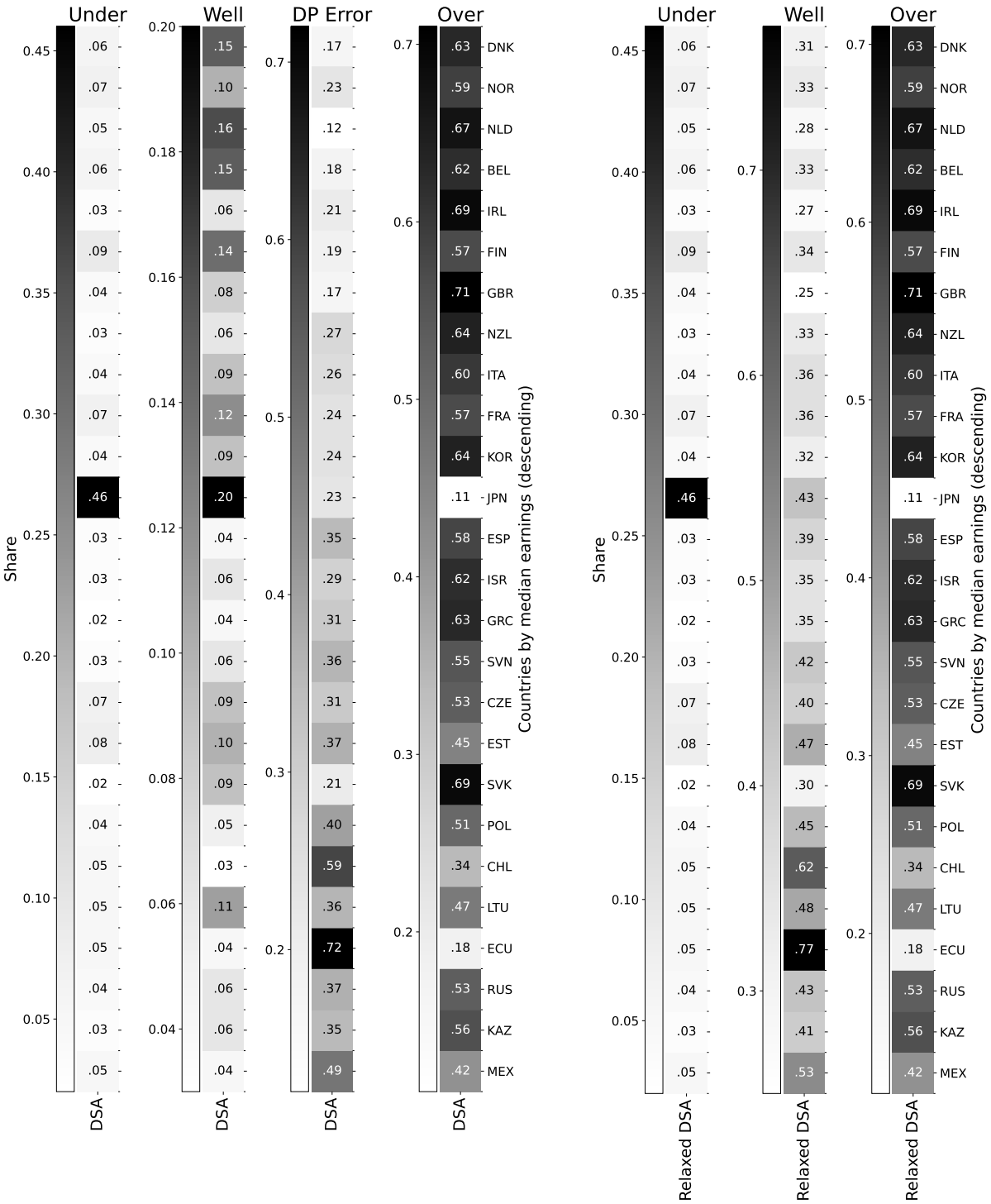
Due to the size, the figure is moved to the next page.

Figure 8: Indirect self-assessment: country-specific shares



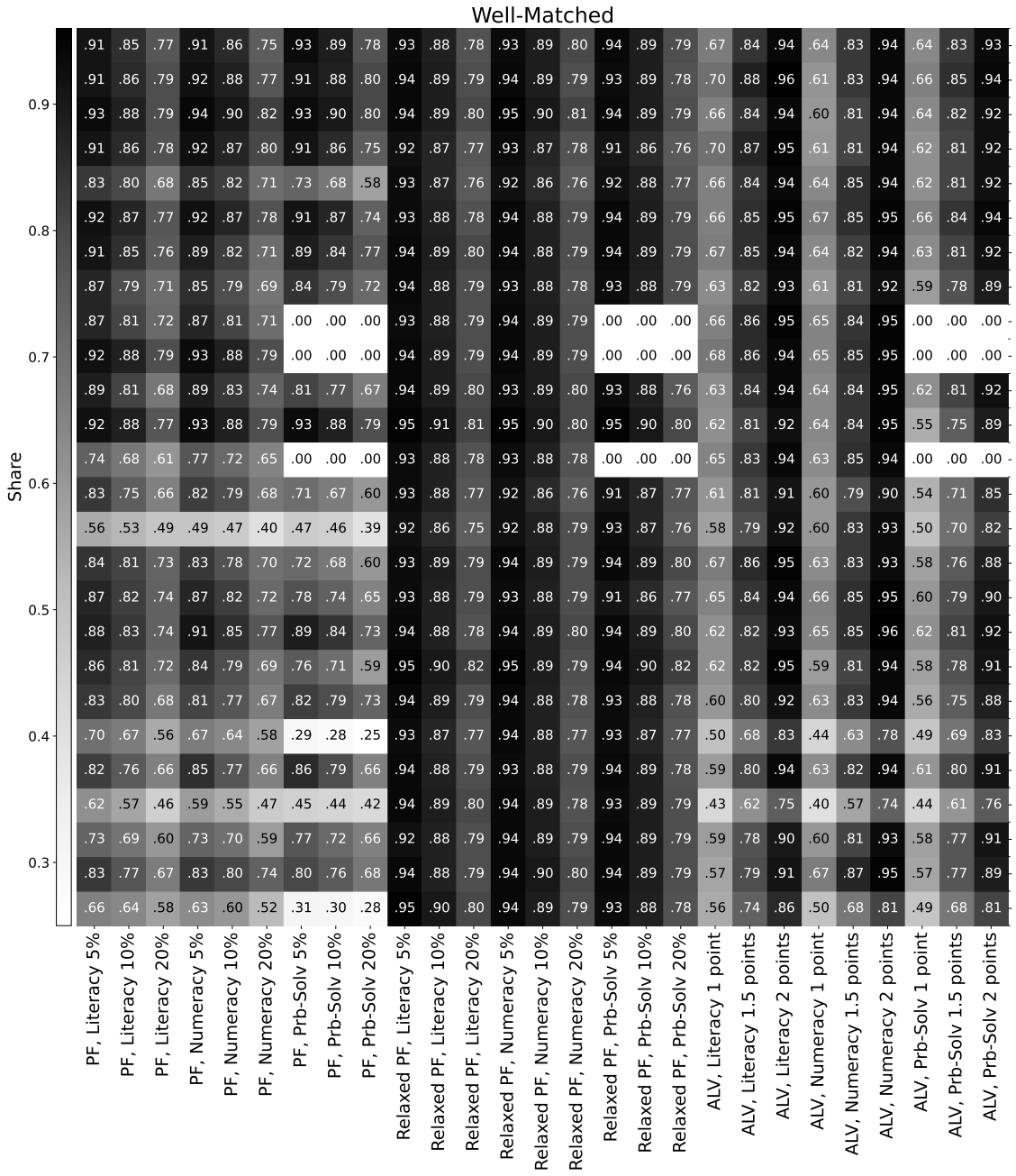
Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD).

Figure 9: Direct self-assessment: country-specific shares



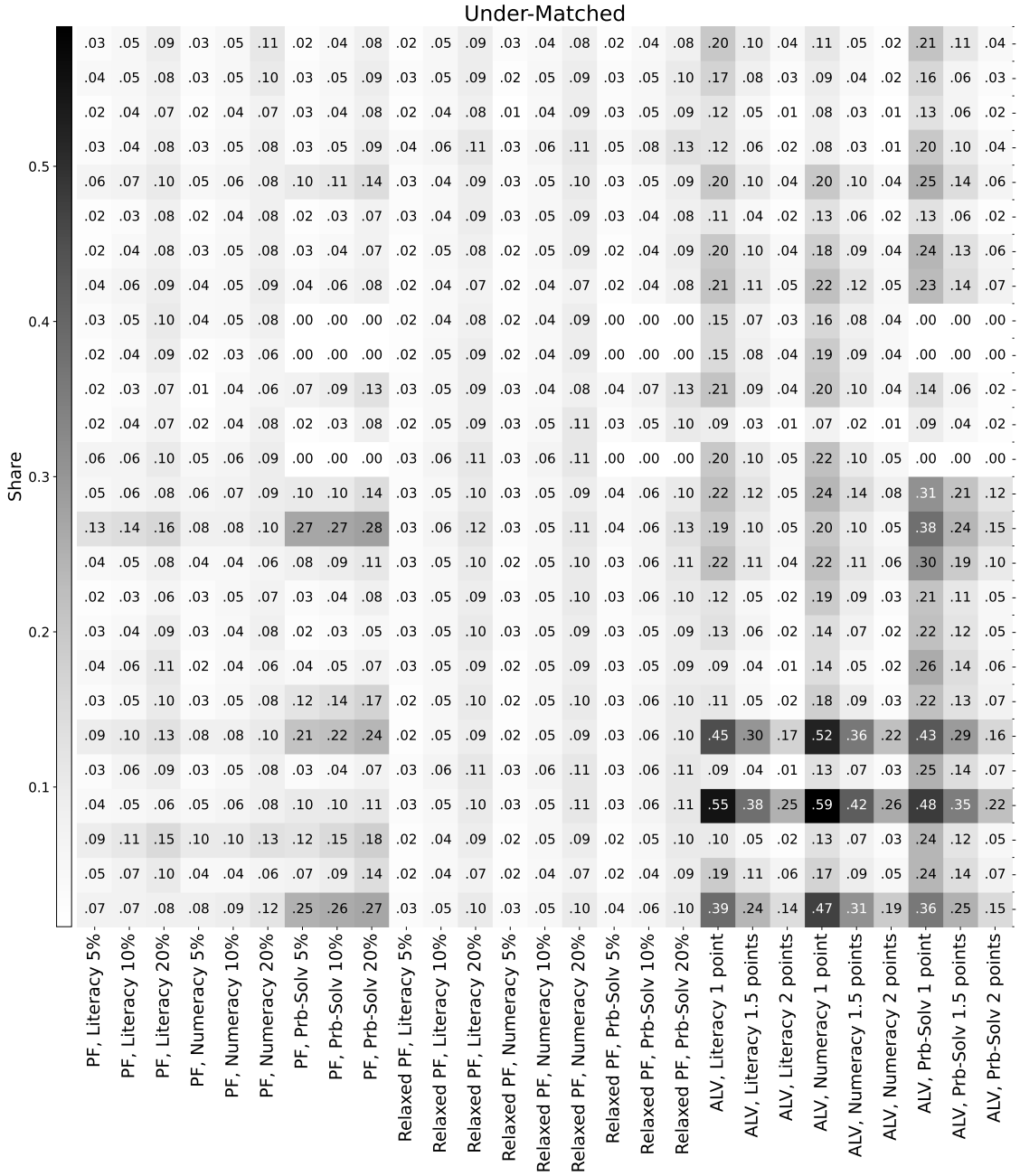
Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD).

Figure 10: Well-skilling: country-specific shares



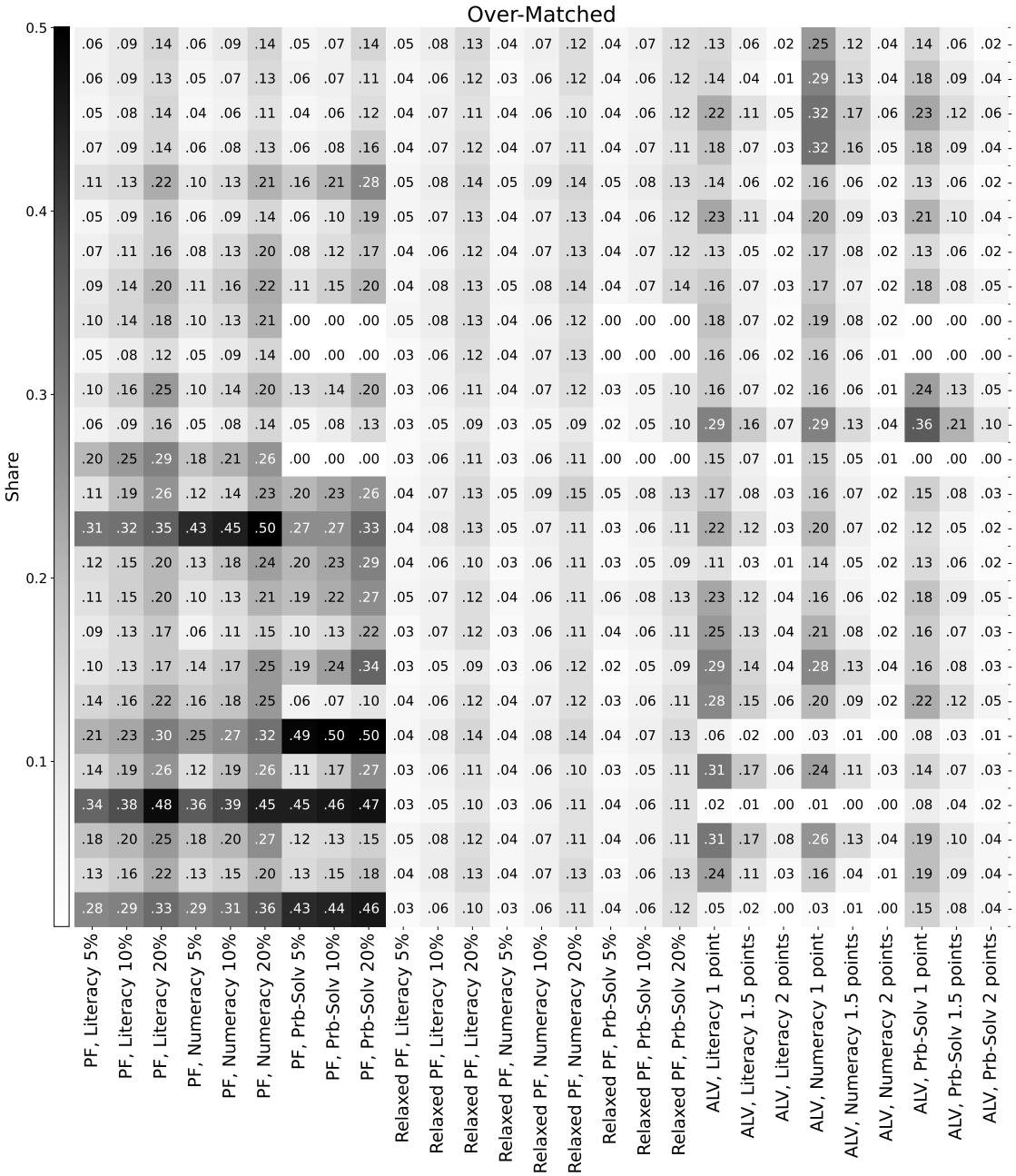
Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD). Problem-solving data is missing for Italy, France and Spain.

Figure 11: Under-skilling: country-specific shares



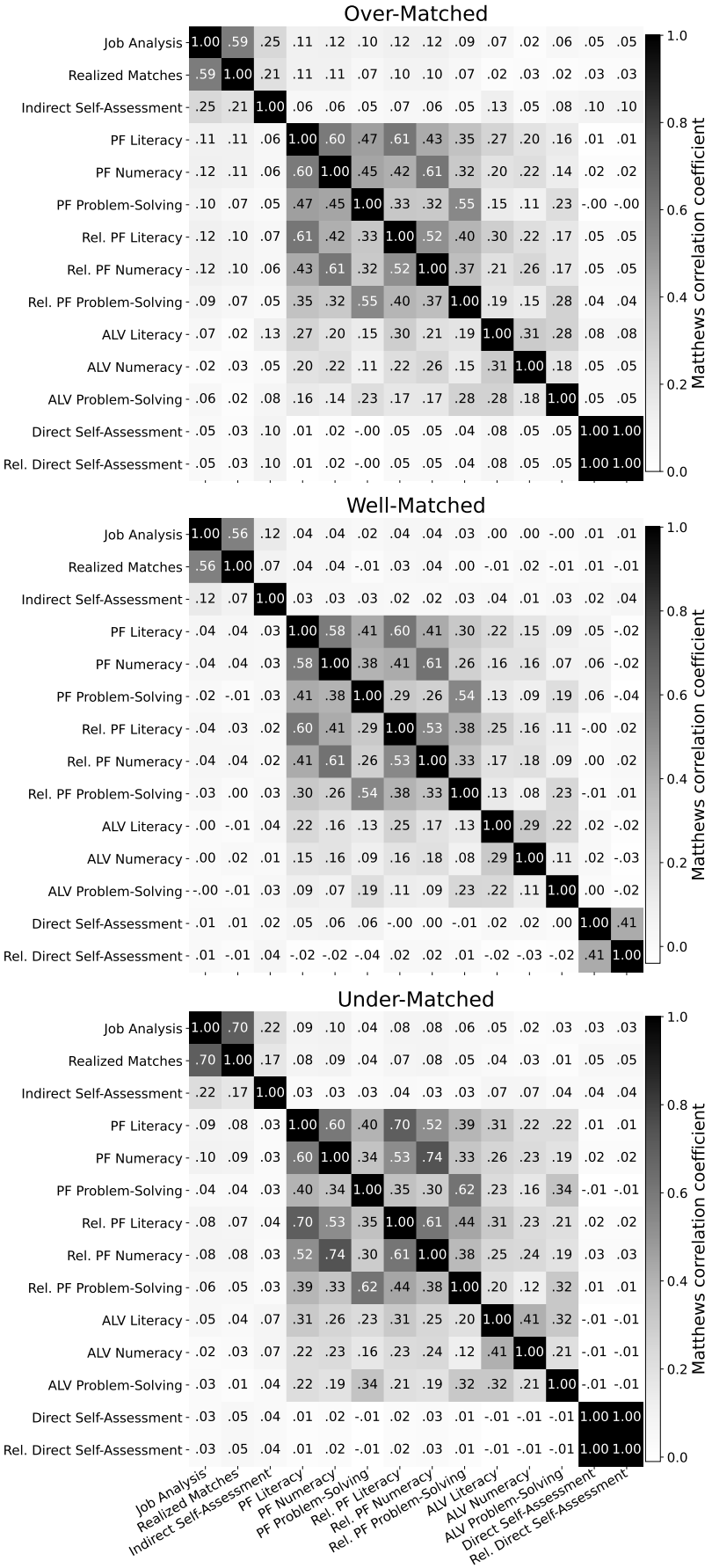
Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD). Problem-solving data is missing for Italy, France and Spain.

Figure 12: Over-skilling: country-specific shares



Notes: Rows are sorted by the country-specific median of hourly earnings including bonuses (PPP corrected USD). Problem-solving data is missing for Italy, France and Spain.

Figure 13: Main measures: correlation analysis



C.1 Heterogeneity in earnings

Figure 14: Wage Gaps: Gender, Age, Migration

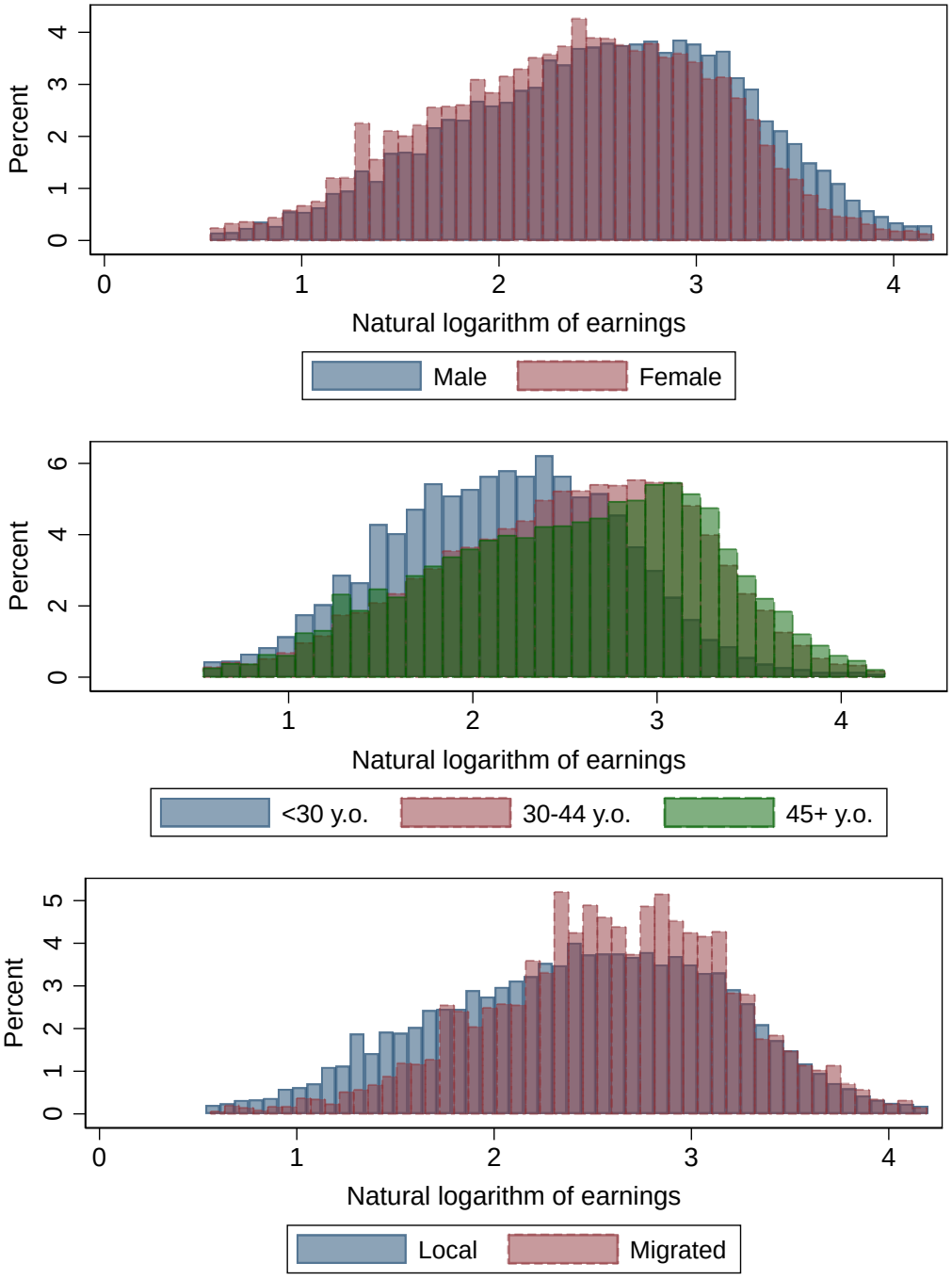
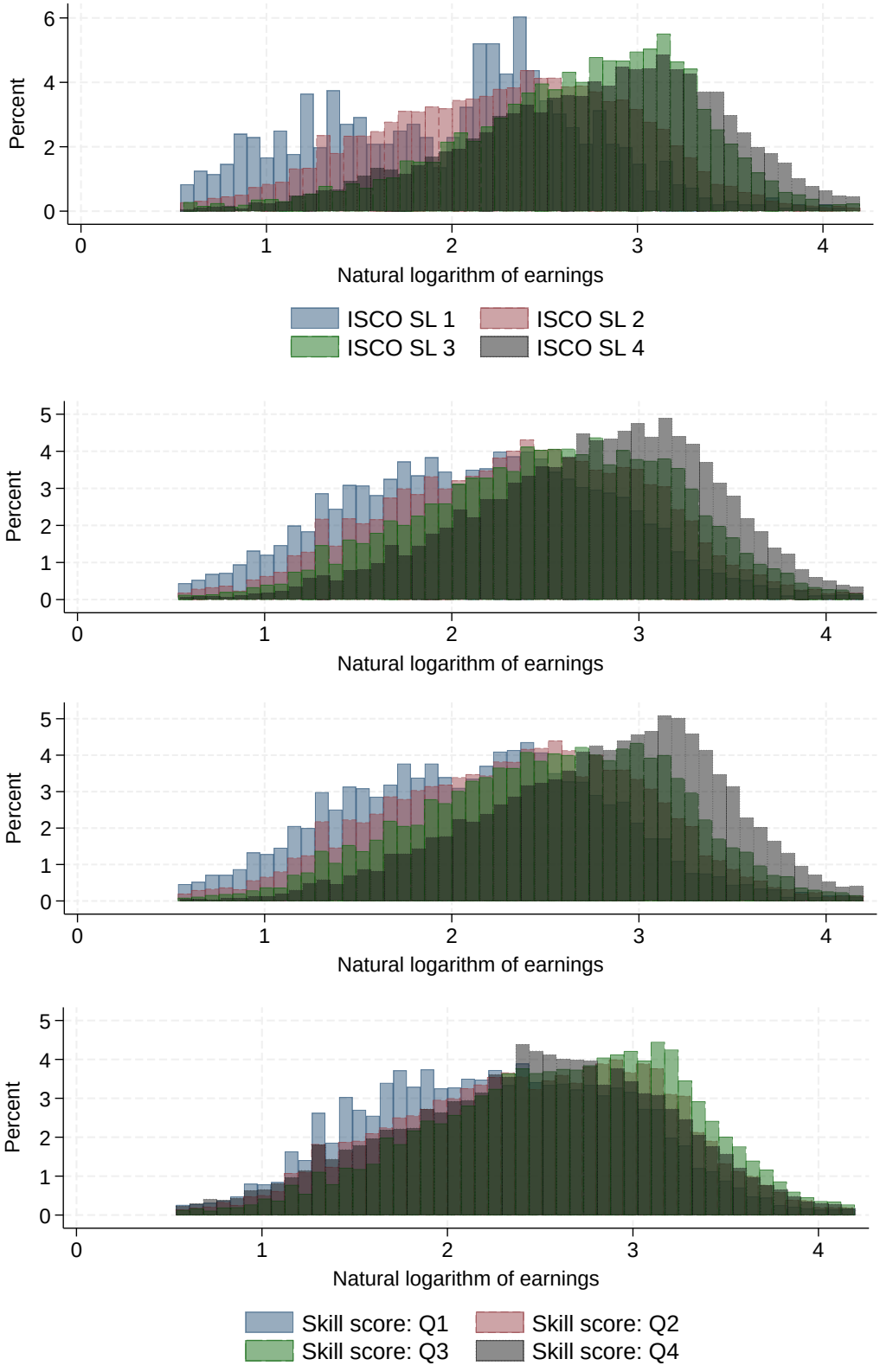
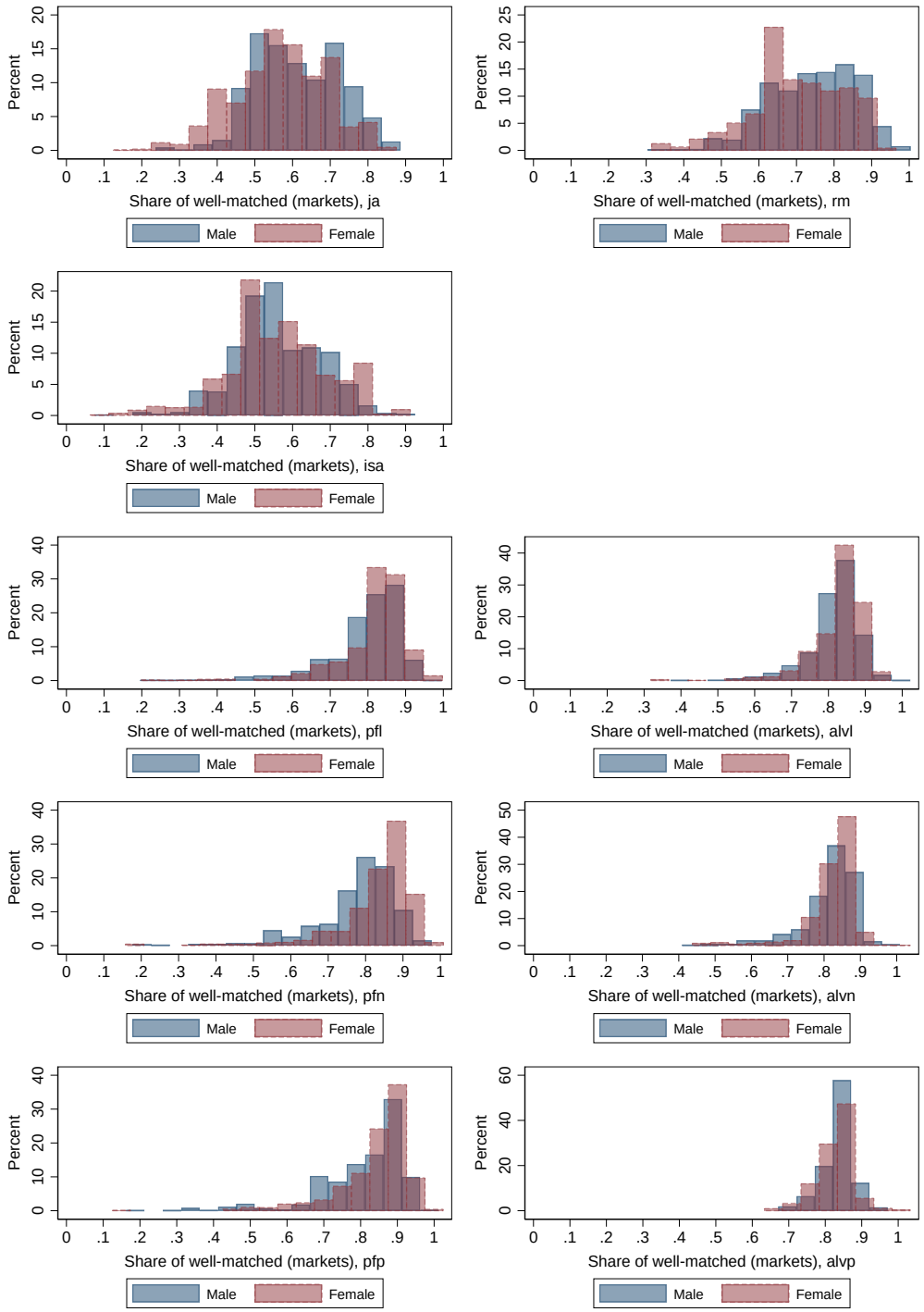


Figure 15: Wage gaps: education, literacy, numeracy, problem-solving



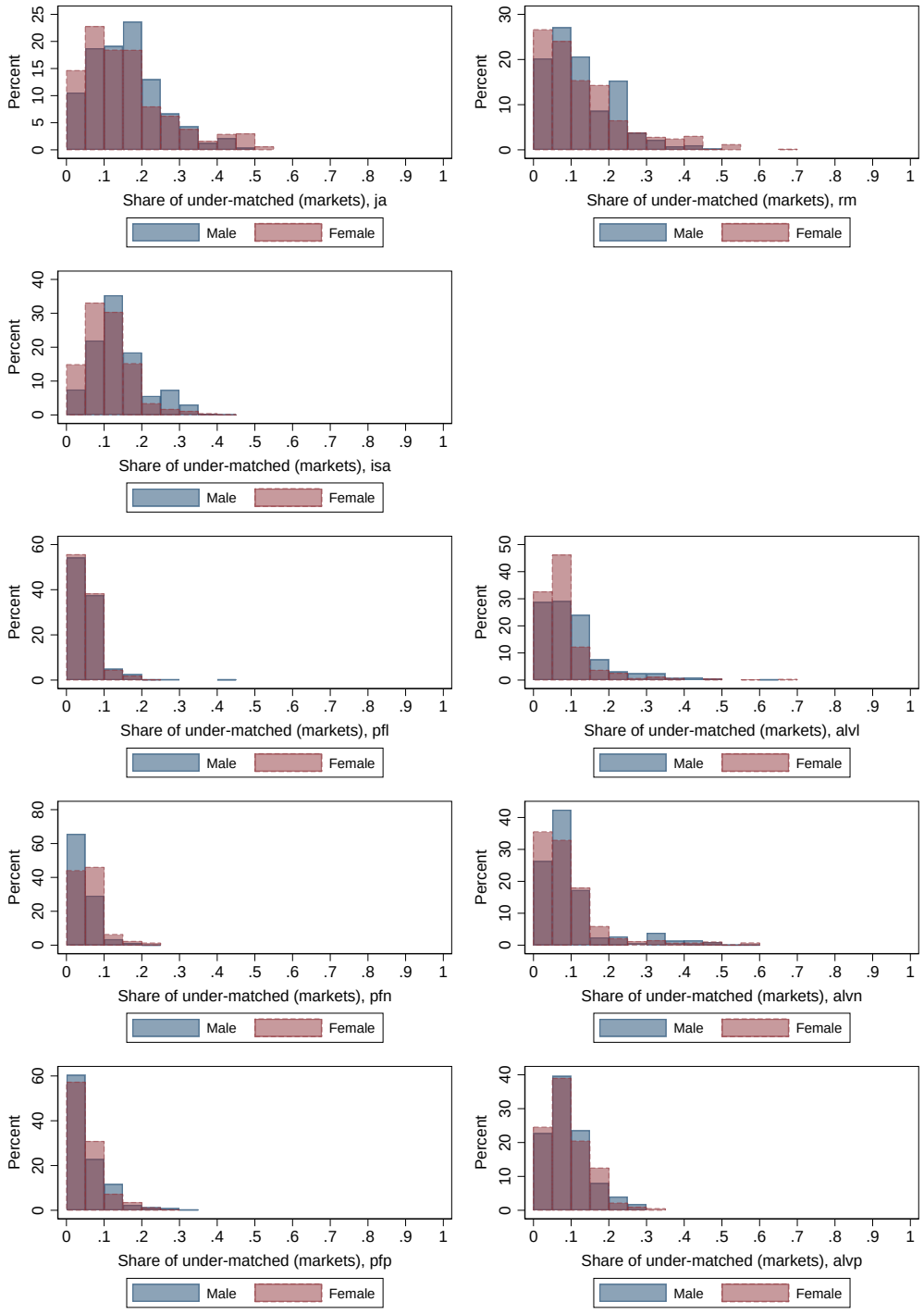
C.2 Heterogeneity in labour mismatch

Figure 16: Shares of well-matched workers by gender



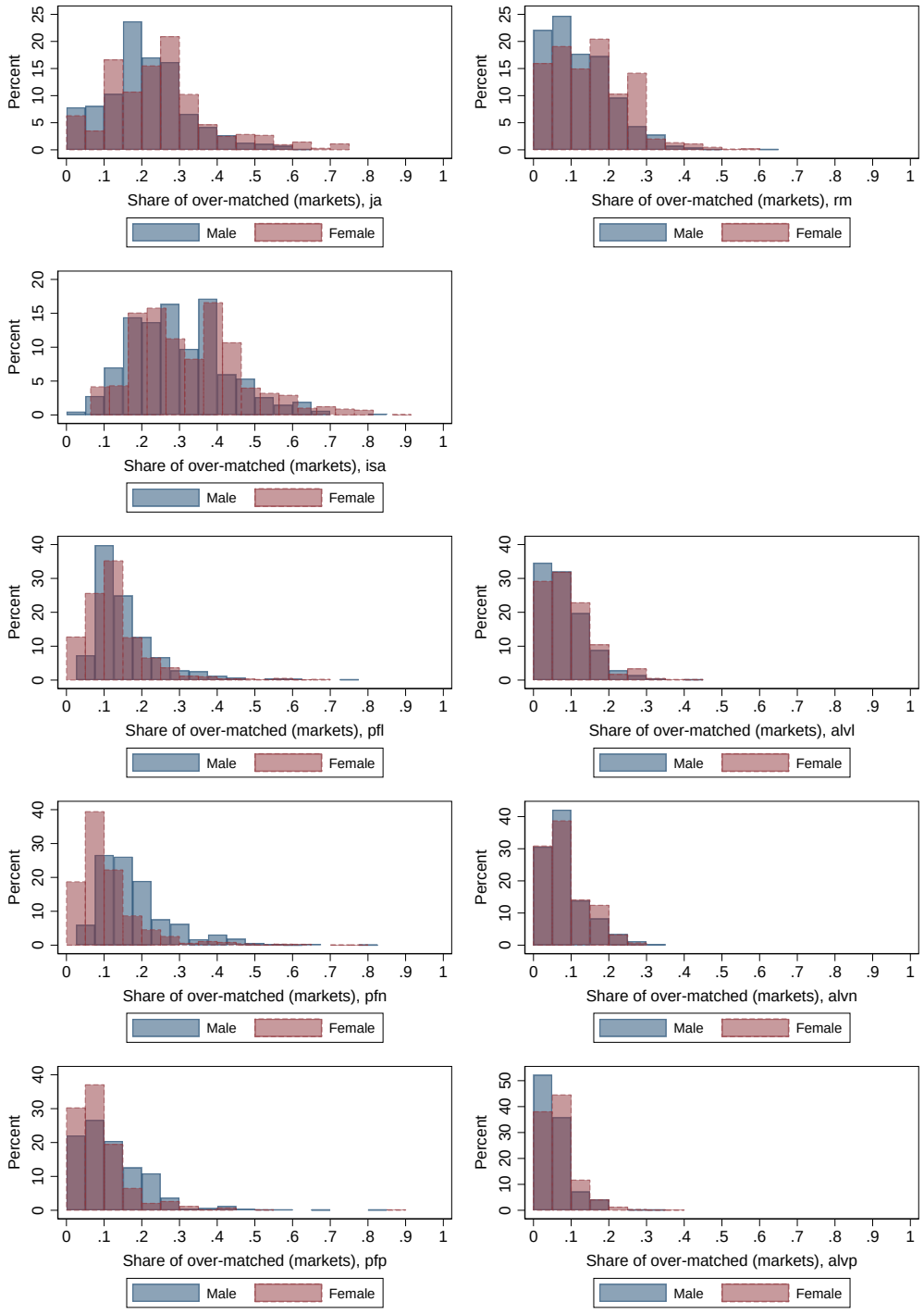
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 17: Shares of under-matched workers by gender



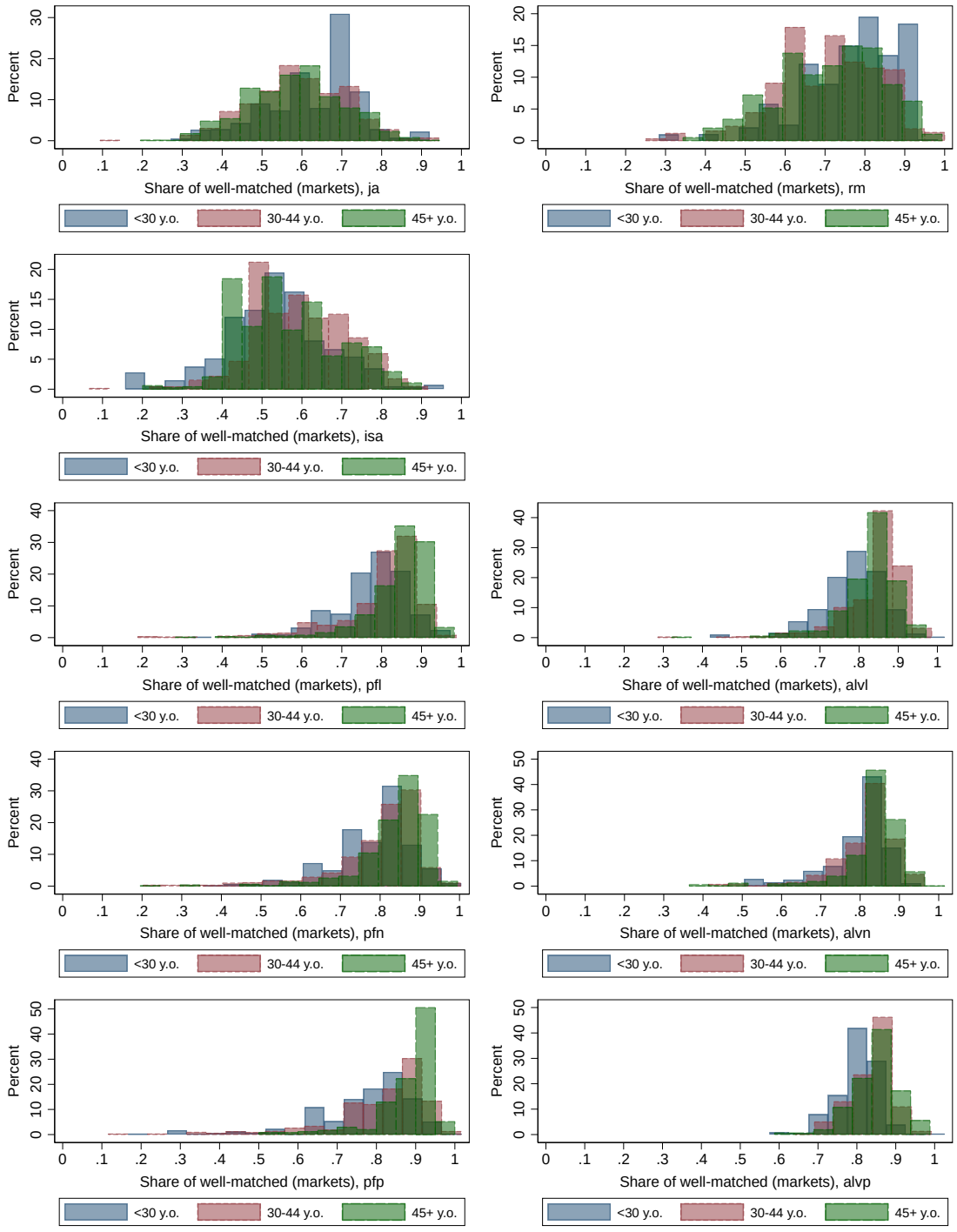
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 18: Shares of over-matched workers by gender



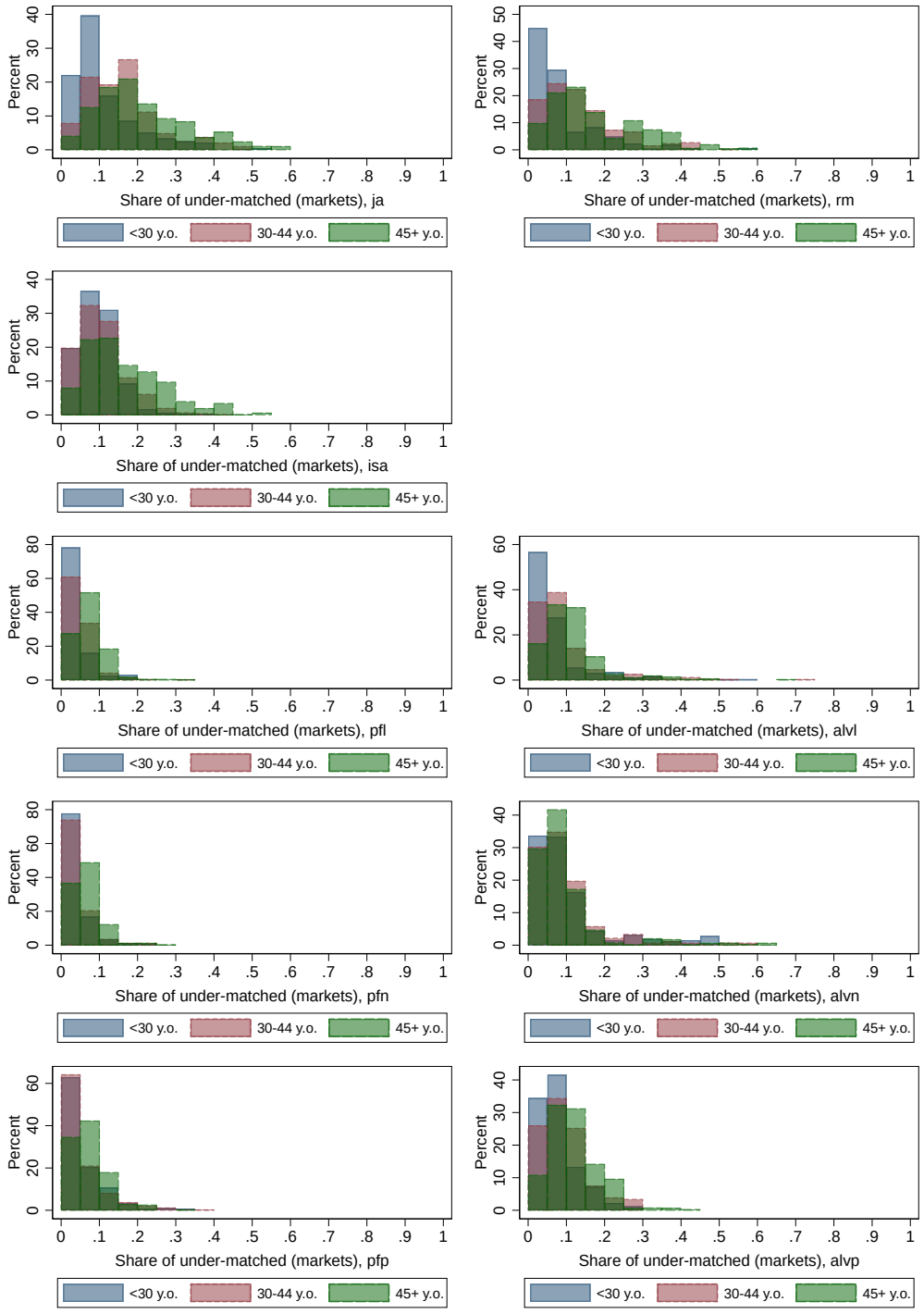
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 19: Shares of well-matched workers by age



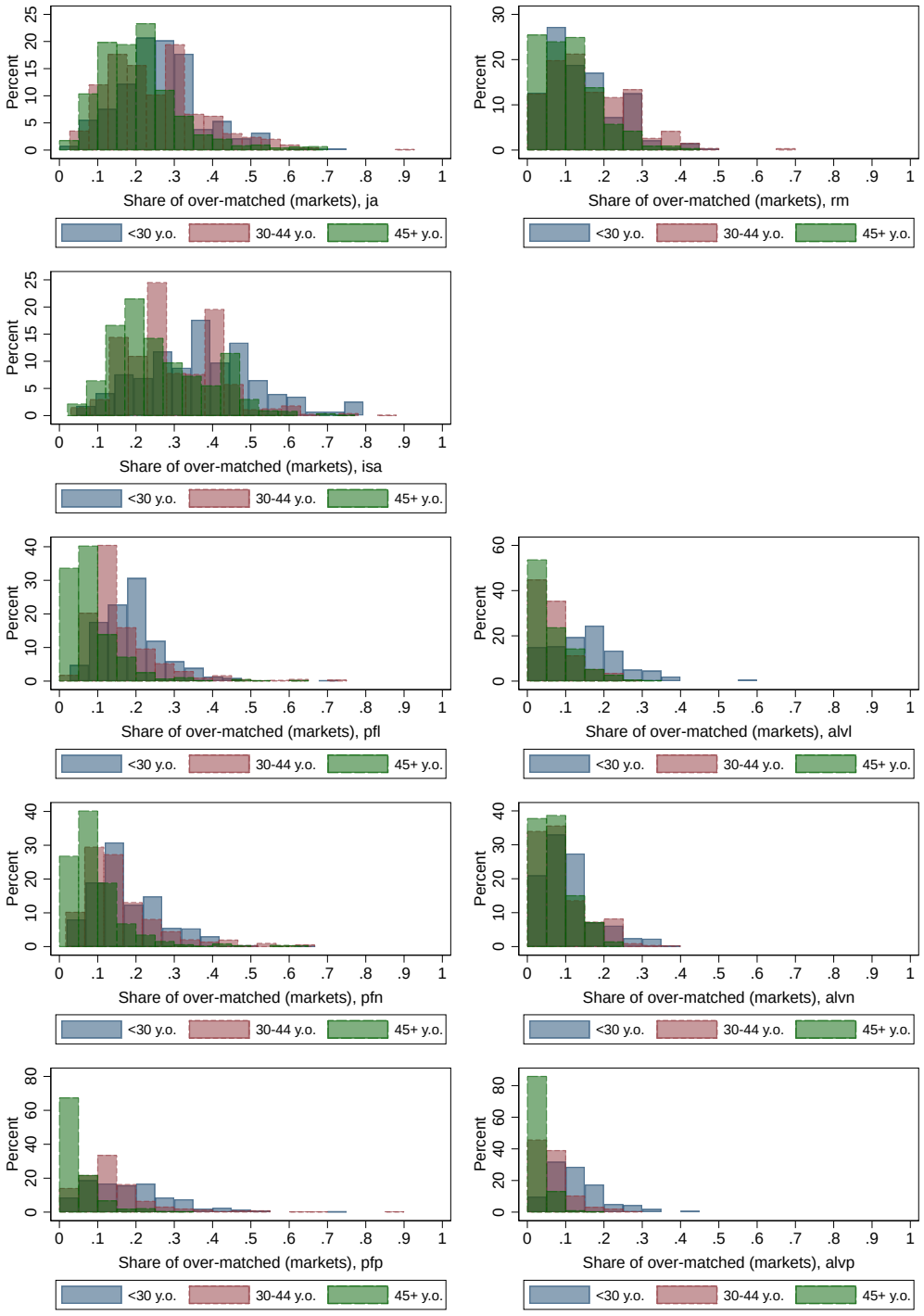
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 20: Shares of under-matched workers by age



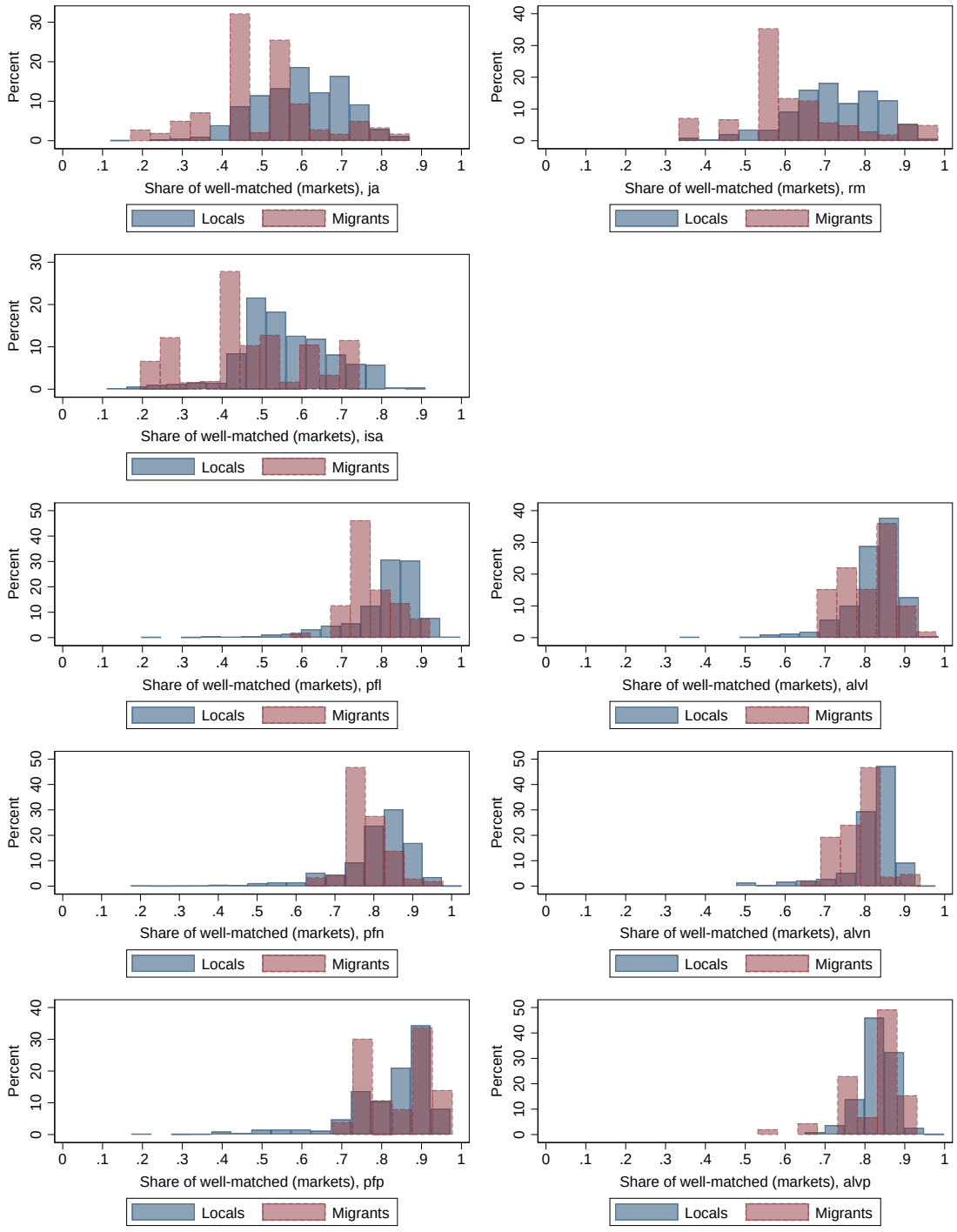
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 21: Shares of over-matched workers by age



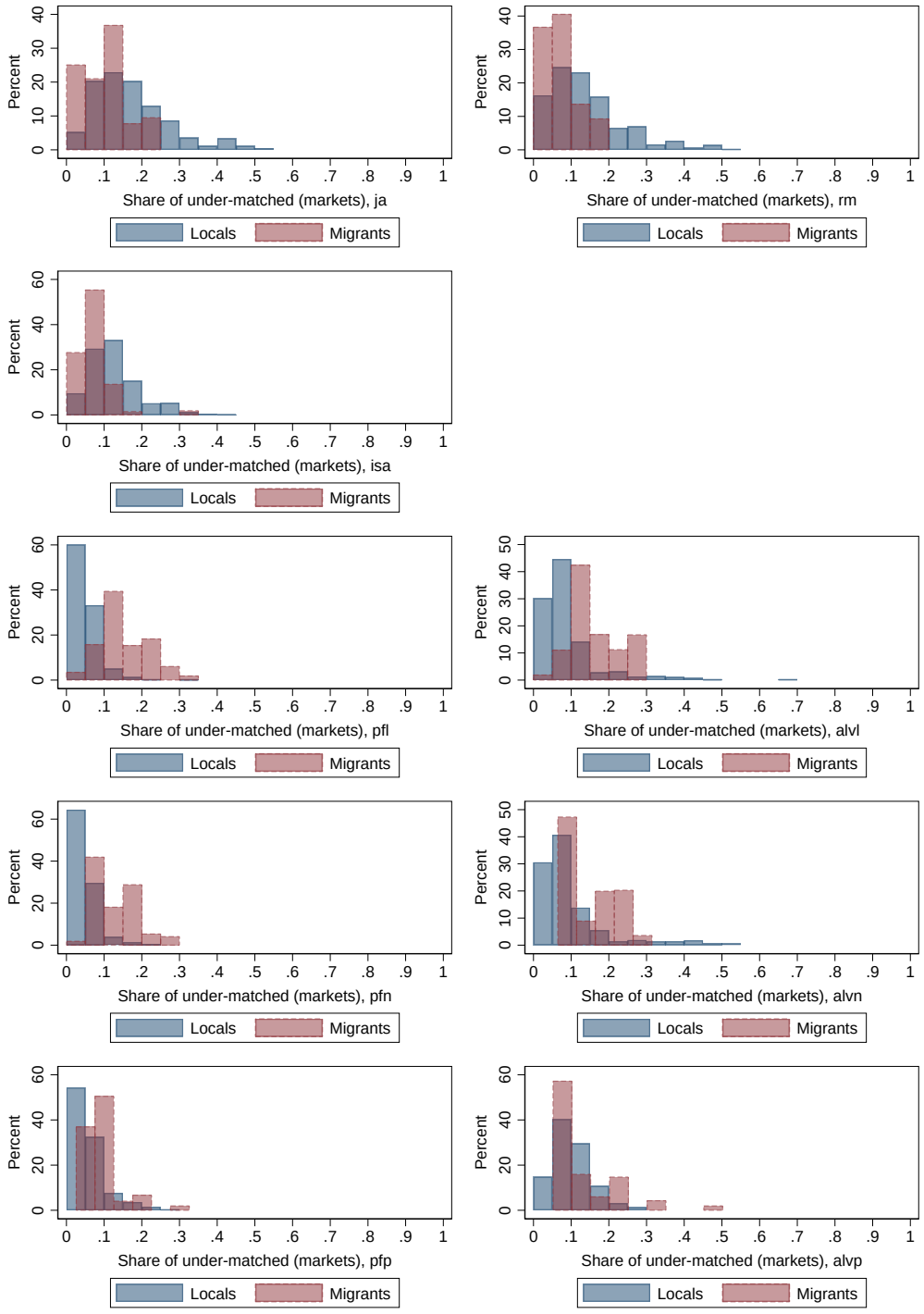
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 22: Shares of well-matched workers by migration



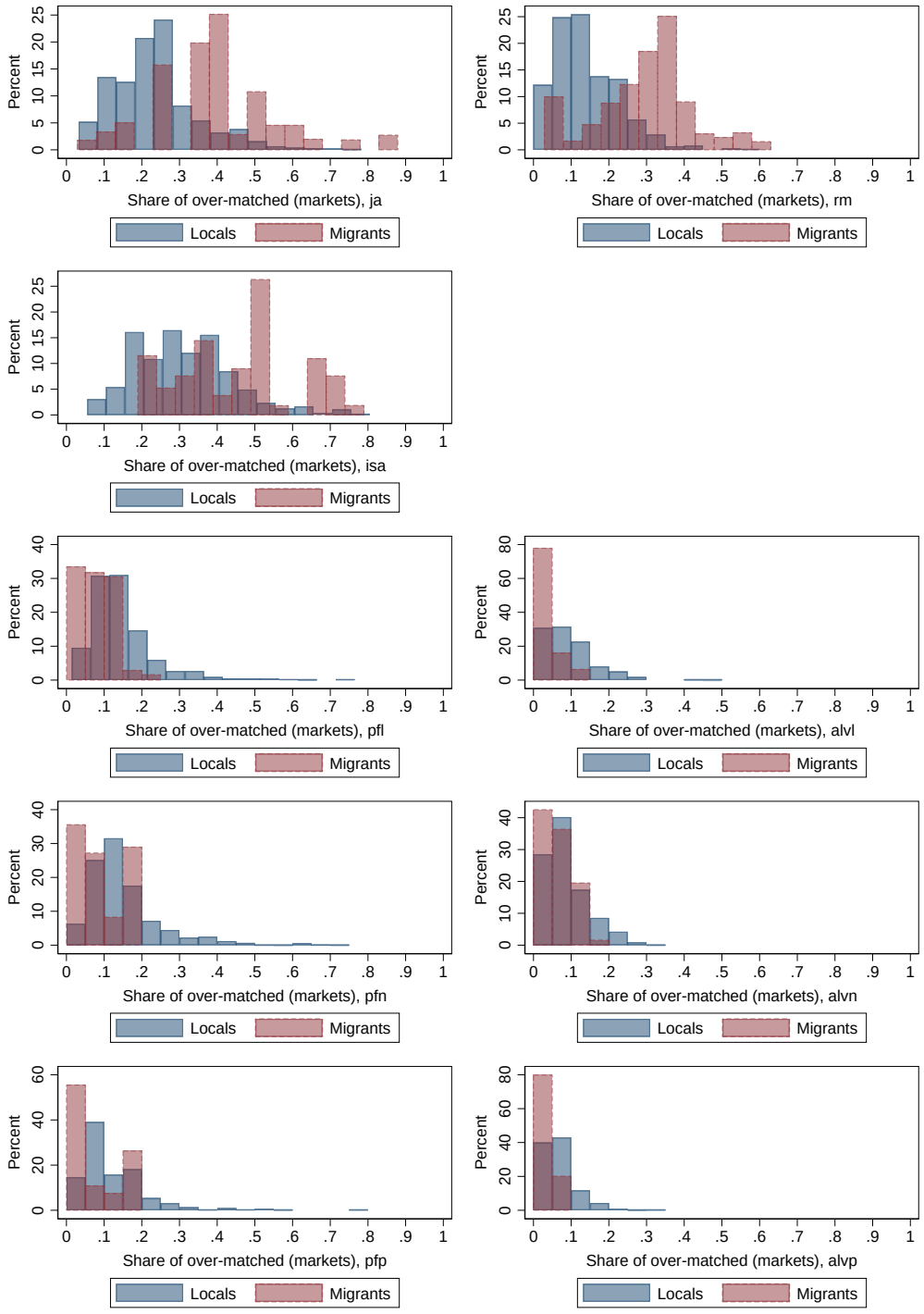
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 23: Shares of under-matched workers by migration



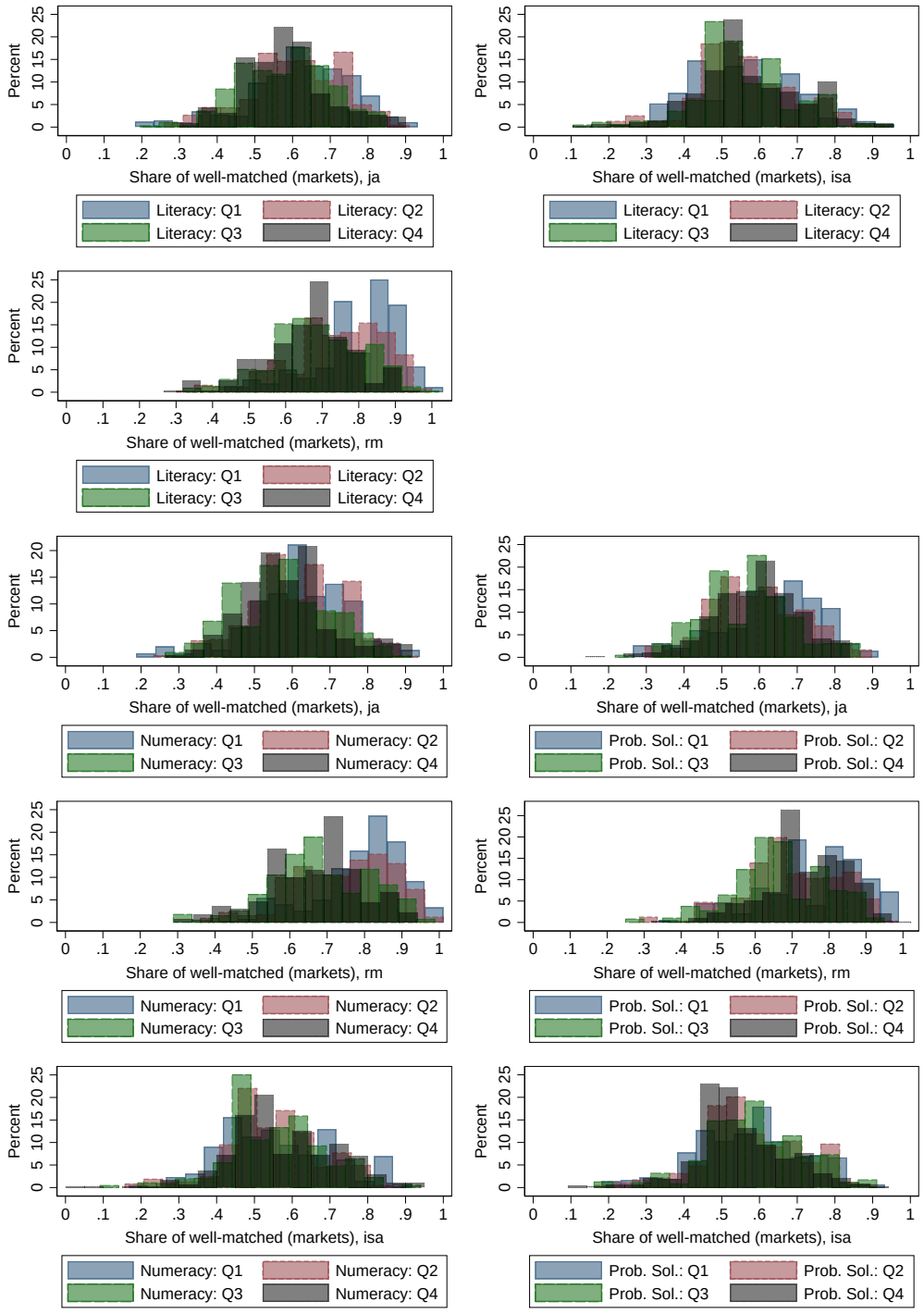
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 24: Shares of over-matched workers by migration



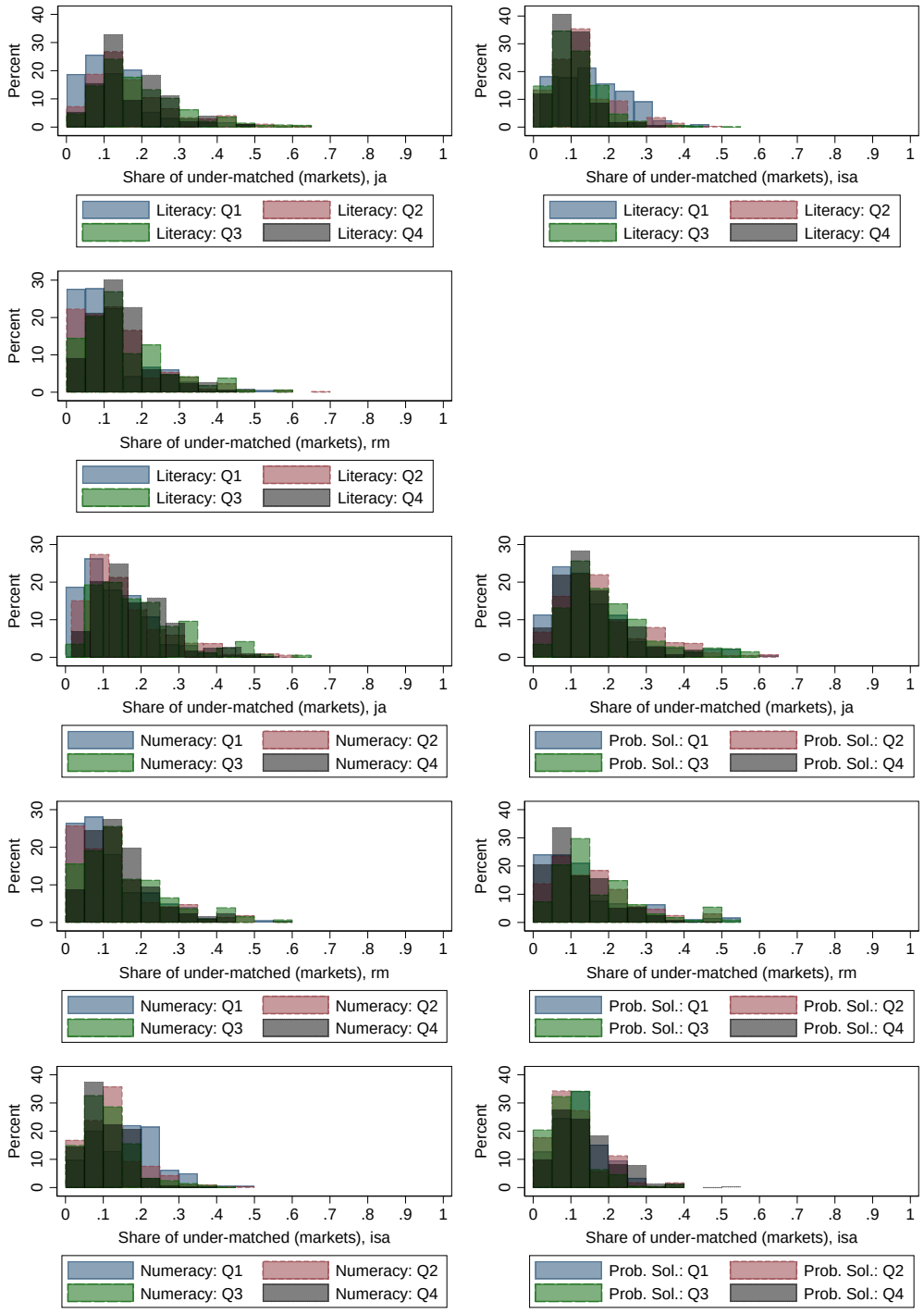
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 25: Shares of well-educated by literacy, numeracy and problem-solving



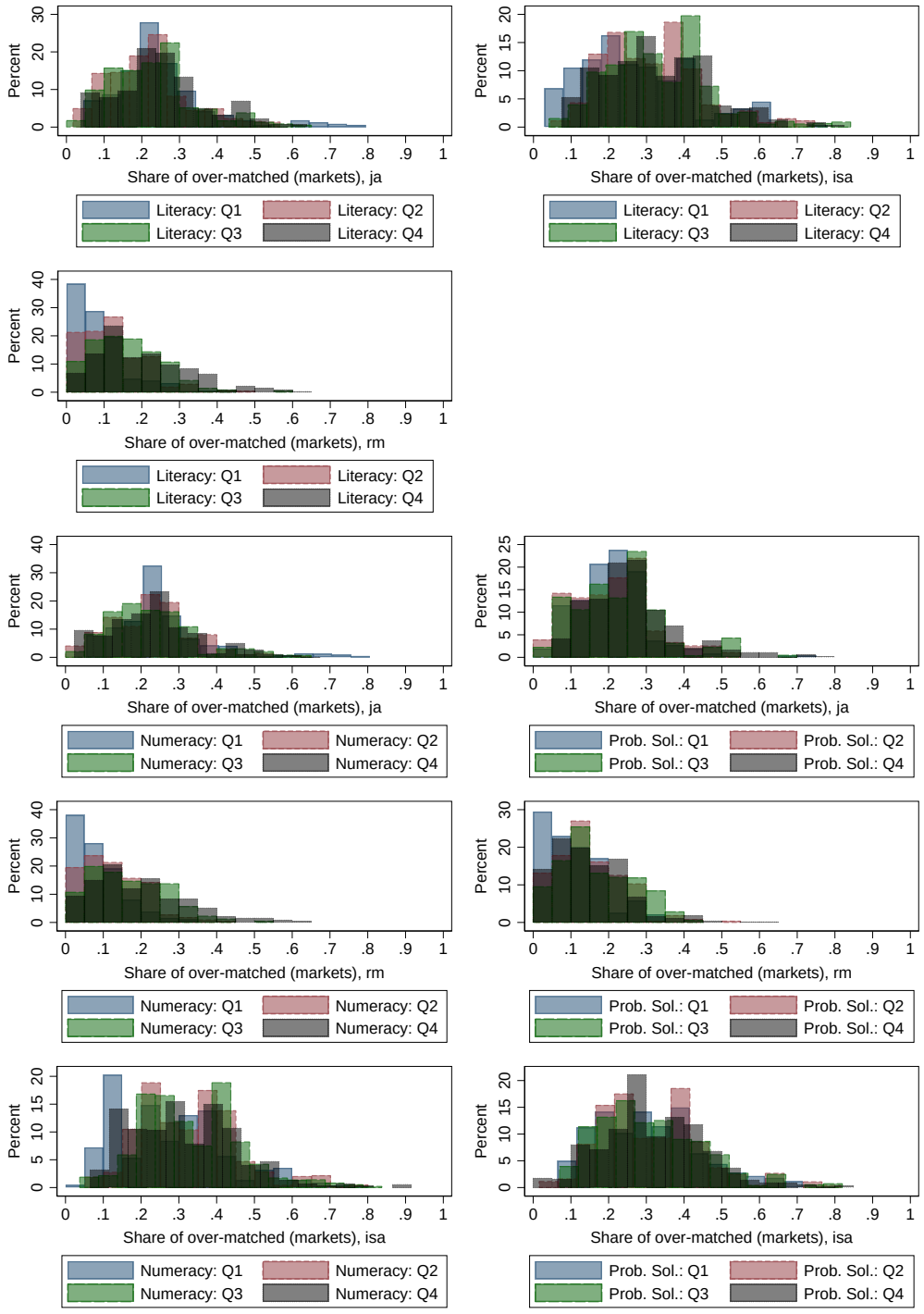
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 26: Shares of under-educated workers by literacy, numeracy and problem-solving



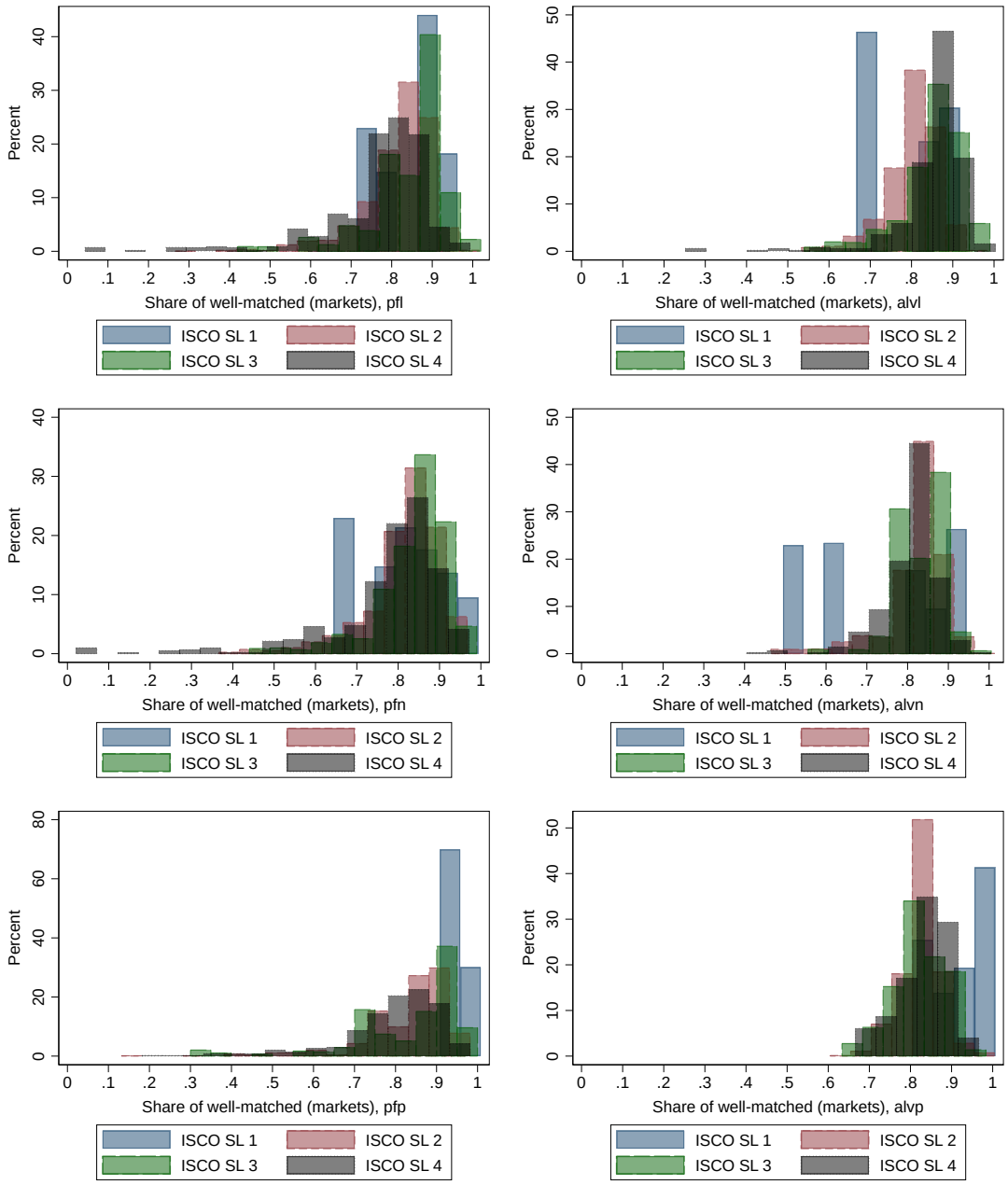
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 27: Shares of over-educated workers by literacy, numeracy and problem-solving



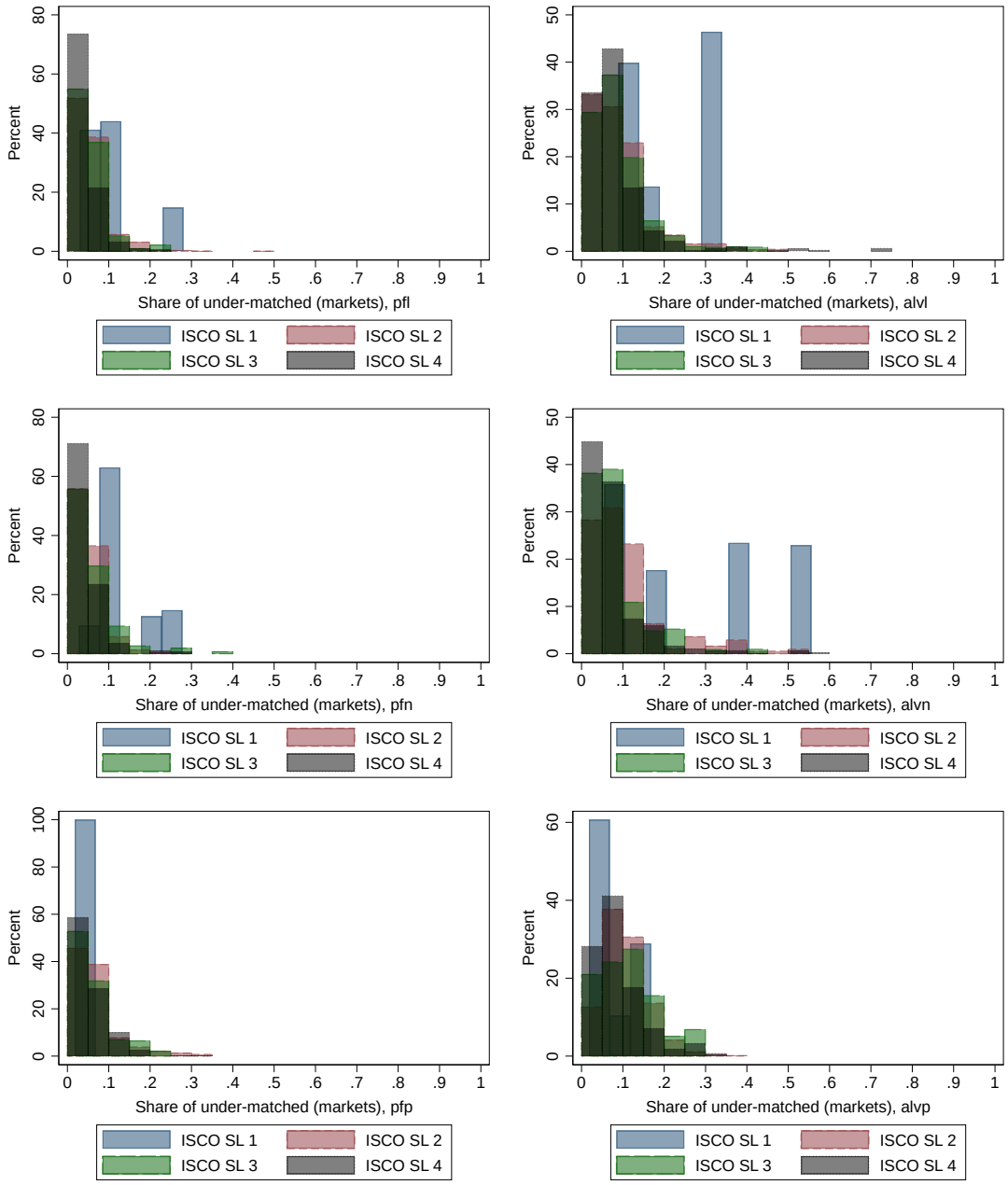
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 28: Shares of well-skilled workers by education



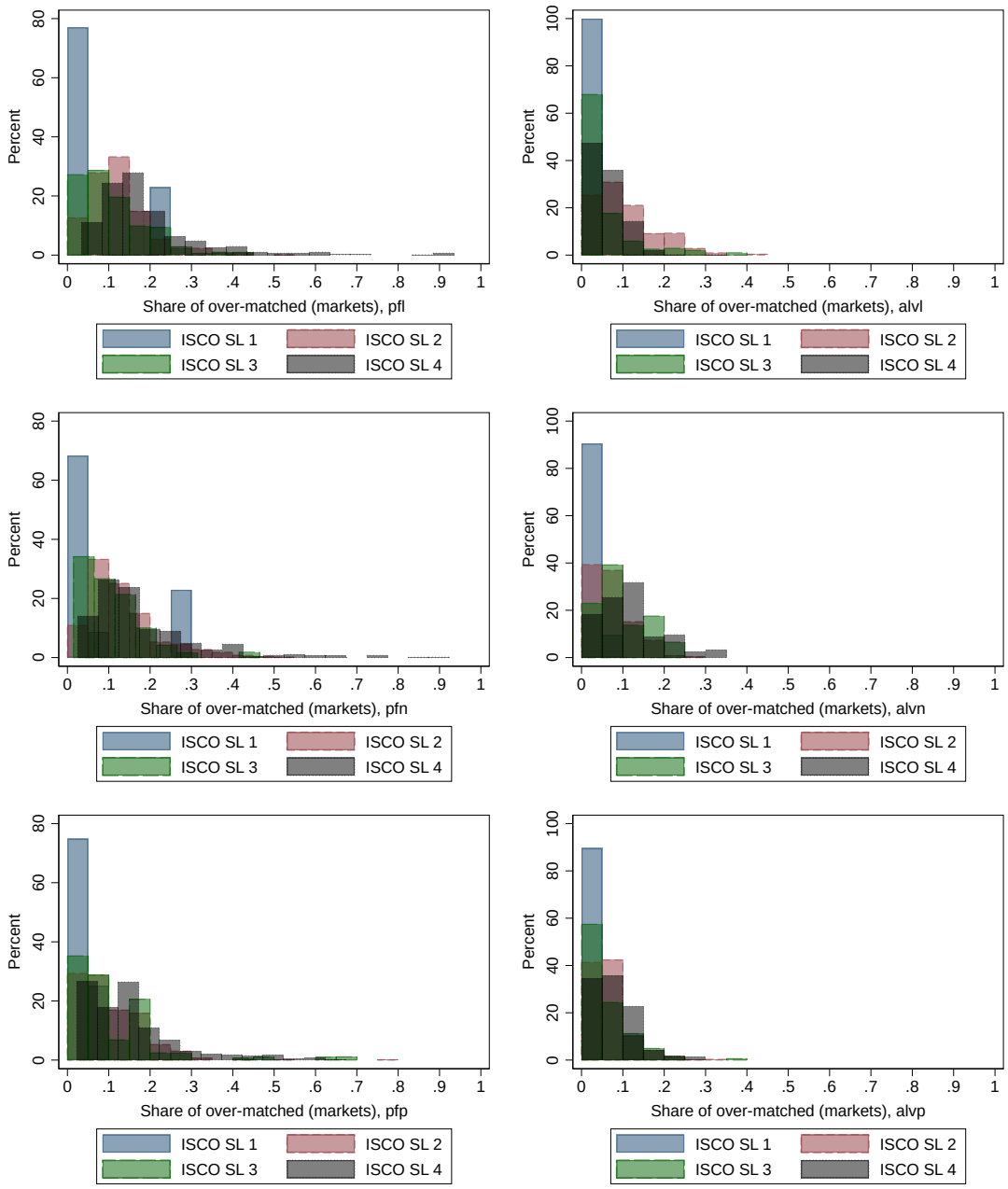
Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 29: Shares of under-skilled workers by education



Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

Figure 30: Shares of over-skilled workers by education



Notes: the shares are calculated at the market level.
Notation: ja – Job Analysis, rm – Realised Matches, isa – Indirect Self Assessment, pfl – Pellizzari-Fichen Literacy, pfn – Pellizzari-Fichen Numeracy, pfp – Pellizzari-Fichen Problem Solving, alvl – Allen-Levels-van-der-Velden Literacy, alvn – Allen-Levels-van-der-Velden Numeracy, alvl – Allen-Levels-van-der-Velden Problem Solving.

D Estimation results

D.1 Lasso model selection

Table 16: Estimation results of Lasso-selected models

	Adaptive Lasso	Lasso, λ_{lopt}	Lasso, λ_{lse}
Female	-0.128	-0.127	-0.124
Age	0.032	0.029	0.015
Age × Age	-0.000	-0.000	-0.000
Tenure	0.007	0.007	0.007
Migrated after 16	-0.065	-0.053	-0.019
Years in country	0.001	0.001	
CHL	-0.649	-0.626	-0.529
CZE	-0.818	-0.799	-0.719
DNK	0.130	0.143	0.195
ECU	-0.982	-0.956	-0.851
FIN	-0.143	-0.106	-0.024
GBR	-0.198	-0.177	-0.096
GRC	-0.671	-0.645	-0.545
IRL	-0.022	0.010	0.072
ISR	-0.474	-0.452	-0.368
JPN	-0.379	-0.360	-0.277
KAZ	-1.258	-1.236	-1.137
KOR	-0.334	-0.313	-0.223
LTU	-1.045	-1.026	-0.942
MEX	-1.033	-1.007	-0.910
NLD	-0.054	-0.036	0.018
NOR	0.081	0.099	0.161
POL	-0.857	-0.838	-0.764
RUS	-1.347	-1.325	-1.230
SVK	-0.918	-0.900	-0.819
SVN	-0.744	-0.724	-0.631
B	0.487	0.431	0.372
C	0.144	0.093	0.063
D	0.209	0.151	0.084
E	0.088	0.027	
F	0.159	0.108	0.074
G	0.043	-0.008	-0.037
H	0.132	0.081	0.048
I	0.032	-0.020	-0.055
J	0.161	0.108	0.073
K	0.217	0.164	0.129
L	0.166	0.107	0.041
M	0.096	0.043	0.006

N	0.057	0.003	-0.010
O	0.114	0.062	0.031
P	0.010	-0.040	-0.062
Q	0.048	-0.001	-0.021
R	0.035	-0.016	-0.040
S	0.031	-0.018	-0.036
T	0.128	0.066	
Industry missing	0.095	0.029	
Education=2	-0.118	-0.043	-0.012
Education=3	-0.079		
Education=4	-0.038	0.055	0.075
Isco required=2	0.027		-0.005
Isco required=3	0.128	0.090	0.075
Isco required=4	0.238	0.185	0.170
Years at school	0.016	0.015	0.013
Years to get job	0.014	0.015	0.018
Not challenged=1	-0.006	-0.004	
Need training=1	-0.007	-0.008	-0.007
Literacy	0.000	0.000	0.000
Numeracy	0.001	0.001	0.001
Problem-Solving	0.001	0.001	0.000
Literacy use	0.054	0.056	0.062
Numeracy use	-0.001	0.001	0.002
Problem-Solving use	0.001	0.002	0.002
ja=0	-0.046	-0.029	
ja=2	0.017	-0.003	-0.007
dsa=0	0.006		
dsa=2	-0.007		
dsa_relaxed=0	-0.006		
dsa_relaxed=2	0.009		
isa_1=2	-0.027	-0.026	-0.021
isa_2=0	0.024	0.017	0.003
isa_2=2	-0.003	-0.001	
isa_3=0	-0.010	-0.001	
isa_3=2	0.001		
isa_4=0	0.081	0.068	0.029
isa_4=2	-0.018	-0.012	
isa_5=0	-0.058	-0.047	
isa_5=2	0.019	0.013	
rm_mean_05=0	-0.005	-0.005	-0.010
rm_mean_05=2	-0.009	-0.000	
rm_mean_1=0	0.029	0.023	
rm_mean_1=2	0.006	0.001	
rm_mean_15=0	-0.034	-0.016	
rm_mean_15=2	-0.002	-0.004	

rm_mode_01=0	-0.154	-0.023	-0.000
rm_mode_01=2	-0.082	-0.010	
rm_mode_1=0	0.121		
rm_mode_1=2	0.029	-0.030	
rm_mode_2=0	0.035	0.025	
rm_mode_2=2	-0.005	-0.013	-0.038
pf_lit_0025=0	-0.006		
pf_lit_0025=2	0.025	0.022	0.011
pf_lit_005=0	0.044	0.033	0.001
pf_lit_005=2	0.022	0.018	0.002
pf_lit_01=0	-0.035	-0.025	
pf_lit_01=2	-0.006	-0.004	
pf_num_0025=0	0.026	0.014	0.001
pf_num_0025=2	0.001		
pf_num_005=0	-0.018	-0.005	
pf_num_005=2	-0.007	-0.005	-0.003
pf_num_01=0	0.005		
pf_num_01=2	0.001	-0.001	-0.005
pf_psl_0025=0	0.009		
pf_psl_0025=2	-0.003	-0.000	-0.001
pf_psl_005=0	-0.014	-0.007	
pf_psl_005=2	0.004		
pf_psl_01=0	-0.008	-0.006	
pf_psl_01=2	-0.020	-0.018	-0.010
pf_lit_0025_relaxed=0	-0.052	-0.041	
pf_lit_0025_relaxed=2	-0.016	-0.008	
pf_lit_005_relaxed=0	0.007	0.001	
pf_lit_005_relaxed=2	0.014	0.007	
pf_lit_01_relaxed=0	0.024	0.021	0.010
pf_lit_01_relaxed=2	-0.009	-0.004	
pf_num_0025_relaxed=0	0.023	0.016	0.006
pf_num_0025_relaxed=2	-0.016	-0.011	
pf_num_005_relaxed=0	-0.009		
pf_num_005_relaxed=2	0.011	0.002	
pf_num_01_relaxed=0	0.014	0.014	0.015
pf_num_01_relaxed=2	-0.003	-0.000	
pf_psl_0025_relaxed=0	0.004		
pf_psl_0025_relaxed=2	-0.042	-0.035	-0.012
pf_psl_005_relaxed=0	-0.006		
pf_psl_005_relaxed=2	0.025	0.016	
pf_psl_01_relaxed=0	0.039	0.035	0.018
pf_psl_01_relaxed=2	-0.006	-0.002	
alv_lit_1=0	0.002		
alv_lit_1=2	-0.018	-0.017	-0.014
alv_lit_15=2	-0.004	-0.002	

alv_lit_2=0	0.008	0.004	
alv_lit_2=2	0.022	0.017	
alv_num_1=0	0.013	0.009	
alv_num_1=2	-0.006	-0.003	
alv_num_15=0	-0.018	-0.013	
alv_num_15=2	-0.000		
alv_num_2=0	0.033	0.027	0.003
alv_num_2=2	-0.019	-0.015	
alv_psl_1=0	-0.006	-0.004	
alv_psl_1=2	-0.034	-0.033	-0.027
alv_psl_15=0	0.002		
alv_psl_15=2	-0.002		
alv_psl_2=0	0.010	0.008	
alv_psl_2=2	0.020	0.015	
Constant	1.135	1.157	1.286
Observations	42922	42922	42922

Notes: The mismatch measure specifications have the following notation.

ja – Job Analysis
rm – Realised Matches: mean-based with 0.5, 1 or 1.5 SDs thresholds or mode-based with 0.1, 1 or 2 SDs thresholds
dsa – Direct Self Assessment: regular or relaxed
isa – Indirect Self Assessment: 1-5 year gaps
pf – Pellizzari-Fichen: regular or relaxed, literacy (lit); numeracy (num) or problem-solving (psl) based; 0.025, 0.05 or 0.1 quantile thresholds
alv – Allen-Levels-van-der-Velden: literacy (lit); numeracy (num) or problem-solving (psl) based; 1, 1.5 or 2 z-score gaps

For the mismatch measure variables, the values of 0 and 2 denote under and over-matched, respectively. The well-matched (the value of 1) is taken as the base category.

D.2 Error components model

Table 17: Mincer Function for Job Analysis

	POLS	Mundlak FE	RE
Under-matched	0.026	0.014	0.03
	[-0.01,0.06]	[-0.00,0.03]	[-0.00,0.03]
Over-matched	-0.051**	-0.056***	-0.056***
	[-0.08,-0.02]	[-0.08,-0.03]	[-0.08,-0.04]
Under-matched (mean)		0.368	
		[-0.06,0.80]	
Over-matched (mean)		-0.175	
		[-0.52,0.17]	
Female	-0.160***	-0.136***	-0.137***
	[-0.21,-0.11]	[-0.16,-0.12]	[-0.16,-0.12]
Age	0.049***	0.049***	0.049***
	[0.04,0.06]	[0.04,0.05]	[0.04,0.06]
Age × Age	-0.000***	-0.001***	-0.001***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.012***	0.009***	0.009***
	[0.01,0.01]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	0.066	-0.044	-0.042
	[-0.03,0.16]	[-0.09,0.01]	[-0.09,0.01]
Years in country	-0.003	0.003*	0.003*
	[-0.01,0.00]	[0.00,0.00]	[0.00,0.00]
Net emigration market share	-5.545***	-5.170***	-6.171***
	[-6.34,-4.75]	[-6.15,-4.19]	[-6.78,-5.56]
Net emigration rate (World Bank)	3.492	11.938*	8.024
	[-12.13,19.11]	[0.56,23.32]	[-5.82,21.87]
ln(Literacy)	0.892***	0.209***	0.223***
	[0.67,1.11]	[0.10,0.32]	[0.12,0.33]
ln(Numeracy)	0.214	0.410***	0.409***
	[-0.00,0.43]	[0.31,0.51]	[0.31,0.51]
ln(Prob. Solv.)	0.381***	0.253***	0.255***
	[0.20,0.56]	[0.18,0.33]	[0.18,0.33]
Constant	-7.153***	-14.654***	-3.784***
	[-7.93,-6.37]	[-18.64,-10.66]	[-4.19,-3.38]
Observations	39931	39931	39931
Markets	253	253	253

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 18: Mincer Function for Realised Matches

	POLS	Mundlak FE	RE
Under-matched	0.046*	0.027**	0.027**
	[0.01,0.09]	[0.01,0.05]	[0.01,0.05]
Over-matched	-0.030	-0.034*	-0.033*
	[-0.07,0.02]	[-0.06,-0.01]	[-0.06,-0.01]
Under-matched (mean)		0.187	
		[-0.24,0.62]	
Over-matched (mean)		0.119	
		[-0.22,0.46]	
Female	-0.160***	-0.138***	-0.138***
	[-0.21,-0.11]	[-0.16,-0.12]	[-0.16,-0.12]
Age	0.050***	0.049***	0.049***
	[0.04,0.06]	[0.04,0.06]	[0.04,0.06]
Age × Age	-0.000***	-0.001***	-0.001***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.012***	0.009***	0.009***
	[0.01,0.01]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	0.061	-0.049	-0.047
	[-0.03,0.15]	[-0.10,0.00]	[-0.10,0.00]
Years in country	-0.003	0.003*	0.003*
	[-0.01,0.00]	[0.00,0.00]	[0.00,0.00]
Net emigration market share	-5.575***	-5.579***	-6.208***
	[-6.36,-4.79]	[-6.47,-4.69]	[-6.81,-5.60]
Net emigration rate (World Bank)	2.425	7.718	7.197
	[-13.21,18.06]	[-4.23,19.67]	[-6.63,21.02]
ln(Literacy)	0.884***	0.208***	0.222***
	[0.66,1.10]	[0.10,0.32]	[0.12,0.33]
ln(Numeracy)	0.225*	0.413***	0.412***
	[0.01,0.44]	[0.31,0.52]	[0.31,0.51]
ln(Prob. Solv.)	0.378***	0.254***	0.256***
	[0.20,0.56]	[0.18,0.33]	[0.18,0.33]
Constant	-7.168***	-14.389***	-3.818***
	[-7.96,-6.38]	[-18.48,-10.30]	[-4.24,-3.40]
Observations	39931	39931	39931
Markets	253	253	253

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 19: Mincer Function for Indirect Self Assessment

	POLS	Mundlak FE	RE
Under-matched	0.057**	-0.017	-0.016
	[0.02,0.09]	[-0.04,0.00]	[-0.03,0.00]
Over-matched	-0.146***	-0.094***	-0.095***
	[-0.17,-0.12]	[-0.11,-0.08]	[-0.11,-0.08]
Under-matched (mean)		1.966***	
		[1.48,2.45]	
Over-matched (mean)		-0.353**	
		[-0.60,-0.10]	
Female	-0.157***	-0.136***	-0.136***
	[-0.20,-0.11]	[-0.16,-0.12]	[-0.16,-0.12]
Age	0.049***	0.048***	0.049***
	[0.04,0.06]	[0.04,0.05]	[0.04,0.05]
Age × Age	-0.000***	-0.001***	-0.001***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.011***	0.009***	0.009***
	[0.01,0.01]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	0.081	-0.044	-0.040
	[-0.01,0.17]	[-0.09,0.00]	[-0.09,0.01]
Years in country	-0.003	0.003*	0.002*
	[-0.01,0.00]	[0.00,0.00]	[0.00,0.00]
Net emigration market share	-5.499***	-4.806***	-6.105***
	[-6.25,-4.75]	[-5.54,-4.07]	[-6.70,-5.51]
Net emigration rate (World Bank)	2.631	13.053**	7.110
	[-12.33,17.60]	[4.82,21.28]	[-6.55,20.77]
ln(Literacy)	0.922***	0.216***	0.234***
	[0.71,1.14]	[0.11,0.32]	[0.13,0.34]
ln(Numeracy)	0.229*	0.417***	0.412***
	[0.02,0.44]	[0.32,0.52]	[0.31,0.51]
ln(Prob. Solv.)	0.346***	0.236***	0.242***
	[0.16,0.53]	[0.16,0.31]	[0.16,0.32]
Constant	-7.159***	-19.594***	-3.766***
	[-7.90,-6.41]	[-21.94,-17.24]	[-4.16,-3.37]
Observations	39259	39259	39259
Markets	253	253	253

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 20: Mincer Function for Pellizzari-Fichen Literacy

	POLS	Mundlak FE	RE
Under-matched	-0.126***	-0.019	-0.020
	[-0.16,-0.09]	[-0.04,0.00]	[-0.04,0.00]
Over-matched	-0.071***	0.054***	0.053***
	[-0.10,-0.04]	[0.03,0.07]	[0.03,0.07]
Under-matched (mean)		-1.685***	
		[-2.50,-0.87]	
Over-matched (mean)		-1.039***	
		[-1.36,-0.72]	
Female	-0.184***	-0.156***	-0.157***
	[-0.23,-0.14]	[-0.17,-0.14]	[-0.17,-0.14]
Age	0.039***	0.039***	0.039***
	[0.03,0.05]	[0.03,0.04]	[0.03,0.04]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.010***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.177***	-0.211***	-0.211***
	[-0.23,-0.12]	[-0.24,-0.19]	[-0.24,-0.19]
Years in country	0.002	0.005***	0.005***
	[-0.00,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-6.122***	-5.692***	-6.576***
	[-7.05,-5.19]	[-6.49,-4.89]	[-7.22,-5.93]
Net emigration rate (World Bank)	0.174	9.856	0.063
	[-17.50,17.85]	[-3.50,23.21]	[-14.54,14.66]
ISCO SL 2	0.294***	0.157***	0.157***
	[0.18,0.41]	[0.11,0.20]	[0.11,0.20]
ISCO SL 3	0.588***	0.347***	0.349***
	[0.45,0.72]	[0.30,0.40]	[0.30,0.40]
ISCO SL 4	0.754***	0.554***	0.556***
	[0.62,0.89]	[0.50,0.61]	[0.50,0.61]
Constant	0.959***	1.571***	1.043***
	[0.78,1.14]	[1.05,2.10]	[0.92,1.17]
Observations	57922	57922	57922
Markets	300	300	300

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 21: Mincer Function for Pellizzari-Fichen Numeracy

	POLS	Mundlak FE	RE
Under-matched	-0.109***	-0.030*	-0.031*
	[-0.15,-0.07]	[-0.05,-0.01]	[-0.06,-0.01]
Over-matched	-0.073***	0.062***	0.060***
	[-0.11,-0.04]	[0.04,0.08]	[0.04,0.08]
Under-matched (mean)		-1.239**	
		[-2.04,-0.44]	
Over-matched (mean)		-0.960***	
		[-1.23,-0.69]	
Female	-0.185***	-0.153***	-0.154***
	[-0.23,-0.14]	[-0.17,-0.14]	[-0.17,-0.14]
Age	0.039***	0.039***	0.039***
	[0.03,0.05]	[0.03,0.04]	[0.03,0.04]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.010***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.180***	-0.210***	-0.210***
	[-0.23,-0.12]	[-0.24,-0.18]	[-0.24,-0.18]
Years in country	0.001	0.005***	0.005***
	[-0.00,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-6.116***	-5.860***	-6.589***
	[-7.05,-5.18]	[-6.63,-5.09]	[-7.23,-5.94]
Net emigration rate (World Bank)	0.306	9.249	-0.134
	[-17.36,17.97]	[-3.97,22.47]	[-14.78,14.52]
ISCO SL 2	0.295***	0.153***	0.154***
	[0.18,0.41]	[0.11,0.20]	[0.11,0.20]
ISCO SL 3	0.590***	0.343***	0.345***
	[0.45,0.72]	[0.29,0.39]	[0.30,0.39]
ISCO SL 4	0.755***	0.550***	0.552***
	[0.62,0.89]	[0.50,0.60]	[0.50,0.60]
Constant	0.957***	1.461***	1.048***
	[0.78,1.14]	[0.92,2.00]	[0.92,1.17]
Observations	57922	57922	57922
Markets	300	300	300

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 22: Mincer Function for Pellizzari-Fichen Problem-Solving

	POLS	Mundlak FE	RE
Under-matched	-0.242***	-0.061***	-0.064***
	[-0.30,-0.19]	[-0.09,-0.04]	[-0.09,-0.04]
Over-matched	-0.116***	0.026*	0.024
	[-0.16,-0.07]	[0.00,0.05]	[-0.00,0.05]
Under-matched (mean)		-0.939***	
		[-1.33,-0.54]	
Over-matched (mean)		-0.402**	
		[-0.69,-0.12]	
Female	-0.206***	-0.161***	-0.161***
	[-0.25,-0.16]	[-0.18,-0.14]	[-0.18,-0.14]
Age	0.046***	0.043***	0.043***
	[0.04,0.05]	[0.04,0.05]	[0.04,0.05]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.010***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.117***	-0.180***	-0.179***
	[-0.18,-0.05]	[-0.21,-0.15]	[-0.21,-0.15]
Years in country	-0.001	0.004***	0.004***
	[-0.00,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-5.838***	-5.922***	-6.768***
	[-6.77,-4.91]	[-6.86,-4.98]	[-7.44,-6.10]
Net emigration rate (World Bank)	0.291	8.706	5.642
	[-17.33,17.91]	[-5.47,22.88]	[-9.56,20.85]
ISCO SL 2	0.142**	0.134***	0.134***
	[0.06,0.23]	[0.07,0.19]	[0.07,0.19]
ISCO SL 3	0.366***	0.296***	0.296***
	[0.27,0.46]	[0.23,0.36]	[0.23,0.36]
ISCO SL 4	0.521***	0.499***	0.499***
	[0.43,0.61]	[0.43,0.56]	[0.44,0.56]
Constant	1.082***	1.287***	1.016***
	[0.93,1.24]	[0.70,1.87]	[0.87,1.16]
Observations	39532	39532	39532
Markets	253	253	253

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 23: Mincer Function for Allen-Levels-Van-der-Velden Literacy

	POLS	Mundlak FE	RE
Under-matched	-0.082***	0.041***	0.040***
	[-0.13,-0.04]	[0.02,0.06]	[0.02,0.06]
Over-matched	-0.082***	-0.079***	-0.079***
	[-0.12,-0.04]	[-0.10,-0.06]	[-0.10,-0.06]
Under-matched (mean)		-1.026***	
		[-1.41,-0.64]	
Over-matched (mean)		-0.932*	
		[-1.70,-0.16]	
Female	-0.182***	-0.155***	-0.156***
	[-0.23,-0.13]	[-0.17,-0.14]	[-0.17,-0.14]
Age	0.038***	0.038***	0.038***
	[0.03,0.04]	[0.03,0.04]	[0.03,0.04]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.010***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.179***	-0.224***	-0.224***
	[-0.24,-0.12]	[-0.25,-0.20]	[-0.25,-0.20]
Years in country	0.001	0.005***	0.005***
	[-0.00,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-6.107***	-6.036***	-6.554***
	[-7.04,-5.17]	[-6.89,-5.18]	[-7.20,-5.91]
Net emigration rate (World Bank)	-0.148	-0.799	0.210
	[-17.91,17.61]	[-14.85,13.25]	[-14.33,14.75]
ISCO SL 2	0.295***	0.166***	0.167***
	[0.19,0.40]	[0.12,0.21]	[0.12,0.21]
ISCO SL 3	0.588***	0.357***	0.358***
	[0.46,0.72]	[0.31,0.41]	[0.31,0.41]
ISCO SL 4	0.748***	0.567***	0.569***
	[0.62,0.88]	[0.52,0.62]	[0.52,0.62]
Constant	0.970***	1.242***	1.073***
	[0.78,1.16]	[0.64,1.84]	[0.95,1.20]
Observations	57889	57889	57889
Markets	300	300	300

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 24: Mincer Function for Allen-Levels-Van-der-Velden Numeracy

	POLS	Mundlak FE	RE
Under-matched	-0.138***	0.014	0.013
	[-0.18,-0.10]	[-0.00,0.03]	[-0.00,0.03]
Over-matched	0.110***	0.006	0.007
	[0.07,0.15]	[-0.01,0.02]	[-0.01,0.02]
Under-matched (mean)		-0.304	
		[-0.70,0.09]	
Over-matched (mean)		1.749***	
		[0.90,2.60]	
Female	-0.181***	-0.157***	-0.158***
	[-0.23,-0.13]	[-0.17,-0.14]	[-0.17,-0.14]
Age	0.039***	0.039***	0.039***
	[0.03,0.05]	[0.03,0.04]	[0.03,0.04]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.010***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.162***	-0.218***	-0.218***
	[-0.22,-0.10]	[-0.24,-0.19]	[-0.24,-0.19]
Years in country	0.001	0.005***	0.005***
	[-0.00,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-6.089***	-6.591***	-6.576***
	[-7.01,-5.17]	[-7.43,-5.75]	[-7.22,-5.93]
Net emigration rate (World Bank)	0.465	3.500	0.334
	[-17.02,17.95]	[-10.10,17.10]	[-14.20,14.87]
ISCO SL 2	0.266***	0.163***	0.164***
	[0.16,0.37]	[0.12,0.21]	[0.12,0.21]
ISCO SL 3	0.555***	0.355***	0.357***
	[0.43,0.68]	[0.30,0.41]	[0.31,0.41]
ISCO SL 4	0.712***	0.566***	0.568***
	[0.59,0.83]	[0.51,0.62]	[0.52,0.62]
Constant	0.959***	1.046***	1.046***
	[0.79,1.13]	[0.53,1.56]	[0.92,1.17]
Observations	57914	57914	57914
Markets	300	300	300

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

Table 25: Mincer Function for Allen-Levels-Van-der-Velden Problem-Solving

	POLS	Mundlak FE	RE
Under-matched	-0.192***	-0.034***	-0.036***
	[-0.24,-0.14]	[-0.05,-0.02]	[-0.05,-0.02]
Over-matched	0.035*	-0.030**	-0.029**
	[0.01,0.06]	[-0.05,-0.01]	[-0.05,-0.01]
Under-matched (mean)		-1.588***	
		[-2.11,-1.06]	
Over-matched (mean)		0.011	
		[-0.74,0.77]	
Female	-0.201***	-0.161***	-0.162***
	[-0.25,-0.15]	[-0.18,-0.14]	[-0.18,-0.15]
Age	0.047***	0.042***	0.043***
	[0.04,0.06]	[0.04,0.05]	[0.04,0.05]
Age × Age	-0.000***	-0.000***	-0.000***
	[-0.00,-0.00]	[-0.00,-0.00]	[-0.00,-0.00]
Tenure	0.013***	0.009***	0.010***
	[0.01,0.02]	[0.01,0.01]	[0.01,0.01]
Migrated after 16	-0.109**	-0.181***	-0.180***
	[-0.18,-0.04]	[-0.21,-0.15]	[-0.21,-0.15]
Years in country	-0.001	0.004***	0.004***
	[-0.01,0.00]	[0.00,0.01]	[0.00,0.01]
Net emigration market share	-5.946***	-5.973***	-6.743***
	[-6.89,-5.00]	[-6.87,-5.07]	[-7.41,-6.08]
Net emigration rate (World Bank)	0.185	10.480	5.474
	[-17.74,18.11]	[-2.25,23.21]	[-9.67,20.62]
ISCO SL 2	0.141**	0.138***	0.139***
	[0.05,0.23]	[0.08,0.20]	[0.08,0.20]
ISCO SL 3	0.362***	0.301***	0.302***
	[0.26,0.46]	[0.24,0.36]	[0.24,0.36]
ISCO SL 4	0.507***	0.506***	0.507***
	[0.42,0.60]	[0.44,0.57]	[0.44,0.57]
Constant	1.023***	1.450***	1.038***
	[0.87,1.18]	[0.80,2.10]	[0.90,1.18]
Observations	39871	39871	39871
Markets	253	253	253

95% confidence intervals in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Mean values of the controls are not displayed.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). Lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, *20*(1), 176–235.

Albrecht, J., & Vroman, S. (2002). A matching model with endogenous skill requirements. *International Economic Review*, *43*(1), 283–305.

Allen, J., Levels, M., & Van der Velden, R. (2013). Skill mismatch and skill use in developed countries: Evidence from the piaac study.

Allen, J., & Van der Velden, R. (2001). Educational mismatches versus skill mismatches: Effects on wages, job satisfaction, and on-the-job search. *Oxford Economic Papers*, *53*(3), 434–452.

Baltagi, B. H. (2008). *Econometric analysis of panel data* (Vol. 4). Springer.

Becker, G. S. (1964). Human capital theory. *Columbia, New York*, *1964*.

Castro, J., Ortega, L., Yamada, G., & Mata, D. (2022). The magnitude and predictors of overeducation and overskilling in latin america: Evidence from PIAAC.

Cohn, E. (1992). The impact of surplus schooling on earnings: Comment. *The Journal of Human Resources*, *27*(4), 679–682.

Desjardins, R., & Rubenson, K. (2011). An analysis of skill mismatch using direct measures of skills.

Di Pietro, G., & Urwin, P. (2006). Education and skills mismatch in the italian graduate labour market. *Applied Economics*, *38*(1), 79–93.

Dolado, J. J., Jansen, M., & Jimeno, J. F. (2009). On-the-job search in a matching model with heterogeneous jobs and workers. *The Economic Journal*, *119*(534), 200–228.

European Commission. (2022). *The gender pay gap situation in the eu*. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/gender-equality/equal-pay/gender-pay-gap-situation-eu_en

Flisi, S., Goglio, V., Meroni, E., Rodrigues, M., & Vera-Toscano, E. (2014). Occupational mismatch in europe: Understanding overeducation and overskilling for policy making. *JRC Science and Policy Report, Luxembourg: Publication Office of the European Union*.

Flisi, S., Goglio, V., Meroni, E. C., Rodrigues, M., & Vera-Toscano, E. (2017). Measuring occupational mismatch: Overeducation and overskill in europe—evidence from PIAAC. *Social Indicators Research*, *131*(3), 1211–1249.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Friedman, M. (1953). Choice, chance, and the personal distribution of income. *Journal of Political Economy*, *61*(4), 277–290.

Gautier, P. A. (2002). Unemployment and search externalities in a model with heterogeneous jobs and workers. *Economica*, *69*(273), 21–40.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American statistical Association*, *70*(350), 320–328.

Gill, A. M., & Solberg, E. J. (1992). Surplus schooling and earnings: A critique. *The Journal of Human Resources*, *27*(4), 683–689.

Green, F., & McIntosh, S. (2007). Is there a genuine under-utilization of skills amongst the over-qualified? *Applied economics*, *39*(4), 427–439.

Green, F., & Zhu, Y. (2010). Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. *Oxford economic papers*, *62*(4), 740–763.

Guvenen, F., Kuruscu, B., Tanaka, S., & Wiczer, D. (2020). Multidimensional skill mismatch. *American Economic Journal: Macroeconomics*, *12*(1), 210–244.

Hartog, J. (2000). Over-education and earnings: Where are we, where should we go? *Economics of Education Review*, *19*(2), 131–147.

Heckman, J. J., Lochner, L., & Todd, P. E. (2003). Fifty years of Mincer earnings regressions.

ILO. (2012). *International standard classification of occupations 2008 (isco-08): Structure, group definitions and correspondence tables*. International Labour Office.

Kiker, B. F., Santos, M. C., & De Oliveira, M. M. (1997). Overeducation and undereducation: Evidence for portugal. *Economics of Education Review*, *16*(2), 111–125.

Koopmans, T. C., & Beckmann, M. (1957). Assignment problems and the location of economic activities. *Econometrica: Journal of the Econometric Society*, 53–76.

Krahn, H., & Lowe, G. S. (1998). *Literacy utilization in canadian workplaces*. Statistics Canada Ottawa.

Leuven, E., & Oosterbeek, H. (2011). Overeducation and mismatch in the labor market. *Handbook of the Economics of Education*, *4*, 283–326.

Lindenlaub, I. (2017). Sorting multidimensional types: Theory and application. *The Review of Economic Studies*, *84*(2), 718–789.

Lise, J., & Postel-Vinay, F. (2020). Multidimensional skills, sorting, and human capital accumulation. *American Economic Review*, *110*(8), 2328–2376.

McGowan, M. A., & Andrews, D. (2015). Labour market mismatch and labour productivity: Evidence from PIAAC data.

McGowan, M. A., & Andrews, D. (2017). Skills mismatch, productivity and policies: Evidence from the second wave of PIAAC.

Mincer, J. (1974). *Schooling, experience, and earnings*. National Bureau of Economic Research.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, *66*(4), 281–302.

Montt, G. (2017). Field-of-study mismatch and overqualification: Labour market correlates and their wage penalty. *IZA Journal of Labor Economics*, *6*(1), 1–20.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Nieto, S., & Ramos, R. (2017). Overeducation, skills and wage penalty: Evidence for spain using PIAAC data. *Social Indicators Research*, *134*(1), 219–236.

OECD. (2012). Survey of adult skills, programme for the international assessment of adult competencies (PIAAC) [data retrieved on 15/03/2023, https://www.oecd.org/skills/piaac].

OECD. (2013). First results from the survey of adult skills.

Oosterbeek, H., & Van Ophem, H. (2000). Schooling choices: Preferences, discount rates, and rates of return. *Empirical Economics*, *25*(1), 15–34.

Pellizzari, M., & Fichen, A. (2017). A new measure of skill mismatch: Theory and evidence from PIAAC. *IZA Journal of Labor Economics*, *6*(1), 1–30.

Pissarides, C. A. (2000). *Equilibrium unemployment theory*. MIT press.

Pivovarova, M., & Powers, J. M. (2022). Do immigrants experience labor market mismatch? new evidence from the us piaac. *Large-scale Assessments in Education*, *10*(1), 1–23.

Pouliakas, K., & Russo, G. (2015). Heterogeneity of skill needs and job complexity: Evidence from the oecd PIAAC survey.

Ricardo, D. (1951). On the principles of political economy and taxation. *Sraffa, London*.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers*, *3*(2), 135–146.

Rumberger, R. W. (1987). The impact of surplus schooling on productivity and earnings. *Journal of Human Resources*, 24–50.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, *31*(2), 831–880.

Shin, D.-H., & Bills, D. (2021). Trends in educational and skill mismatch in the united states. *Social sciences*, *10*(10), 395.

Sicherman, N., & Galor, O. (1990). A theory of career mobility. *Journal of Political Economy*, *98*(1), 169–192.

Spence, M. S. (1973). Job market signalling. *Quarterly Journal of Economics*, *90*, 225–243.

Thurow, L. (1975). *Generating inequality*. Basic Books.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Tsay, R., Chung, C., & Yeh, H. (2005). The influence of occupational mismatch on earnings in taiwan: A comparison of the standard deviation approach and the assessment approach. *Journal of Population Studies*, *30*(1), 65–95.

Verdugo, R. R., & Verdugo, N. T. (1989). The impact of surplus schooling on earnings: Some additional findings. *Journal of Human Resources*, 629–643.

Verdugo, R. R., & Verdugo, N. T. (1992). Surplus schooling and earnings: Reply to Cohn and to Gill and Solberg. *The Journal of Human Resources*, *27*(4), 690–695.

Verhaest, D., & Omey, E. (2006). The impact of overeducation and its measurement. *Social Indicators Research*, *77*, 419–448.

World Bank. (2023). World development indicators: Adjusted net national income per capita (current US$); population, total [data retrieved on 10/05/2023, https://databank.worldbank.org/source/world-development-indicators].

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.