Online Appendix Accompanying
"Firm Heterogeneity in Skill Returns"

Michael J. Böhm, Khalil Esmkhani, and Giovanni Gallipoli

NOVEMBER 2023

Contents

A Data and Samples Construction

A.1 Data

Base sample. The main data source for our analysis is the *Longitudinal Integrated Database for Health Insurance and Labor Market Studies* (LISA) by Statistics Sweden (SCB). LISA contains employment information (such as employment status, organization and workplace identifiers, industry and, from 2001, occupation), tax records (including labor and capital income) and demographic information (such as age, education) for all individuals 16 years of age and older domiciled in Sweden. LISA starts in 1990, with the most recent data including 2017.

Our measure of earning returns is annual labor income from the employer with highest recorded earnings. This is available for all workers, not top-coded, and includes end-of-year bonuses and performance pay. LISA reports a unique identifier for each individual's "company of employment", a so-called organization number, as well as a workplace identifier, which is the combination of organization number, employment location, and industry. To be consistent with the earning measure, and with the firm literature (see, among others, Card et al., 2013), we use the workplace with the highest earnings in a given year as the worker's "firm".

We keep workers dependently employed in the private nonprimary sector who earn above the *Prisbasbelopp* (the minimum amount of earnings that qualifies for the earnings-related part of the public pension system; see also Edin and Fredriksson, 2000). In 2008, the *Prisbasbelopp* was 41,000 kr or approximately 6,200 USD. We drop all observations with incomplete data (missing test scores, age, or workplace) and restrict the sample to 20–60 year old males. This process results in a sample of approximately 1 million unique workers, 26 thousand firms, and 6.6 million worker $\times$ year observations for the main sample period of 1999–2008.[1] Column (1) of Table A.1 reports summary statistics for the resulting sample.

Measures of cognitive and noncognitive traits. A strength of our data source is that we have access to extensive and consistent measures of workers' cognitive and noncognitive attributes. This information comes from military enlistment tests, which were mandatory for Swedish males before 2007 and typically taken between age 18 and 19. In the early 2000s, Sweden started requiring progressively fewer males to do military service. The service was abolished in 2010. Before 2007, however, all males were required to take the military enlistment tests and test scores are available for almost 90 percent of males born up to the 1980s (e.g., see Figure A.1 in Böhm et al., 2023).

One might worry that certain individuals could deliberately perform badly on these tests to avoid military service. There are, however, several pieces of evidence suggesting this was not a major problem. In particular, we emphasize that employers routinely put considerable weight on military service performance and anecdotal evidence suggests that some positions – like being an officer in the navy – were important for the networks individuals would obtain; a substantial fraction of individuals working in influential positions within Swedish society went through these

[1] We also document results for two alternative periods, 1990–1999 and 2008–2017.

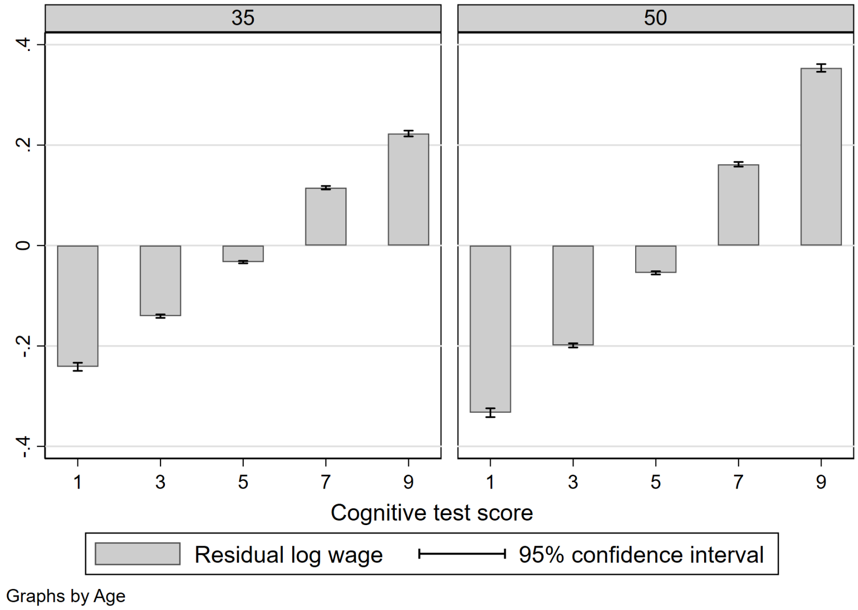
military service assignments. Consistent with these observations, and perhaps more importantly, military test scores have been shown to significantly predict future earnings at long intervals after the tests, as well as other labor market outcomes such as managerial positions and incidence of unemployment (see, e.g., Lindqvist and Vestman, 2011).

The enlistment process for military service spans two days and evaluates a person's medical and physical status as well as cognitive and mental abilities. We use the tests of cognitive and noncognitive ability, which are well established in the labor economics literature, for our analysis. The test of cognitive ability consists of four different parts (logic, verbal, spatial, and technical comprehension), each of which is constructed from 40 questions. These are aggregated into an overall score. The test is a rich measure of general competence and intelligence and it has a stronger fluid IQ component than the American AFQT, which focuses more on crystallized IQ. The aggregate cognitive score ranges from the integer value 1 (lowest) to 9 (highest), according to a STANINE (standard nine) scale that approximates a Normal distribution with a mean of 5 and standard deviation of 2 (meaning that a gap of two scores covers a standard deviation).

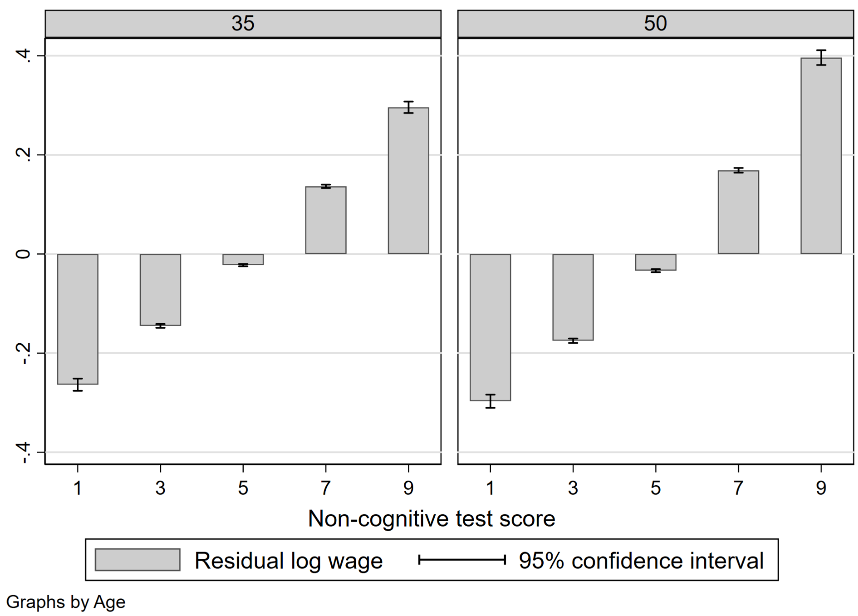
Noncognitive ability is assessed through a 25-minute semi-structured interview by a certified psychologist. Individuals are graded on, among others, their willingness to assume responsibility, independence, outgoing character, persistence, emotional stability, and power of initiative (Swedish National Service Administration, referenced by, among others, Lindqvist and Vestman, 2011). The psychologist weighs these components together and assigns an overall noncognitive score on a STANINE scale. Lindqvist and Vestman (2011), on p. 108f, discuss in detail how the noncognitive score is related to, and different from, other measures often used in the literature on personality and labor market outcomes. Rather than assessing a unique trait, the noncognitive score assesses the ability to function in a demanding environment (military combat). Previous work provides robust evidence that these traits are also rewarded in the labor market.[2]

Test scores and later life outcomes. An important advantage of the military test scores is that they allow for a professional standardized measurement of different ability dimensions over a large population. Military enlistment scores are by design exogenous and predetermined with respect to individuals' career choices. Although cognitive and noncognitive ability are not fixed, they are hard for individuals to manipulate after late childhood or early adulthood (Hansen et al., 2004; Heckman et al., 2006). Crucially, as we show in Figure A.1, the tests are strongly associated to labor market outcomes and accurately predict earnings several decades later. Figure A.1 compares the earnings of workers with different STANINE scores in our sample (residualized using full age $\times$ year dummy interactions) and documents highly significant differences at ages 35 and 50, across both cognitive and noncognitive competencies. These plots emphasize the lasting informational content of the tests and their relevance for long term labor market outcomes. Strong significance at long lags is not always the case with ability tests in survey data and is

[2]Individuals who score sufficiently high on the cognitive test are also evaluated for leadership ability, again on a STANINE scale. The leadership score is meant to capture the suitability to become an officer. Since leadership is only assessed for a subset of individuals, we focus on cognitive and noncognitive ability in our analysis. Noncognitive ability and leadership ability are also highly correlated; in our data the correlation is above 0.8, while the correlation of cognitive and noncognitive is 0.3.



(A) Cognitives



(B) Noncognitives

Figure A.1: Average Earnings of Males at Age 35 and 50, by Test Score Group.

Notes: Earnings for different test score ranks {1,3,5,7,9}; values are residualized using full age × year dummy interactions. Sample period: 1990–2017. 95% confidence intervals indicated by brackets.

partly due to the fine-grained and homogeneous nature of the procedures used to elicit different attributes, resulting in measures that can be mapped into earnings for the whole population of interest over its working life cycle.[3]

A.2 Requirements for Estimating High-Dimensional Models

Estimation requires restrictions on the data samples that guarantee that distinct firm effects can be recovered without biases. These restrictions depend on the empirical approach adopted.

Connected sets. The final sample must consist of firms that are connected to each other through worker mobility. This entails working with a connected component of the firm–worker graph (Abowd et al., 2002; Bonhomme et al., 2023). Distinct connected sets may exist within a large sample of employment matches and empirical analyses often focus on the largest set (or 'maximally connected subgraph'). When considering different skill levels (say high and low cognitive skills) the requirement is that we use a set which is connected for each skill level ("dual" or "double" connected in Card et al., 2016; Kline et al., 2020, respectively). As we show below, the connectedness restrictions become less stringent when observations are defined at the level of firm clusters rather than individual firms.

Limited mobility bias. While connectedness leads to unbiased identification of model parameters, researchers are usually interested in variance components. These may be biased if sampling errors in parameter estimates enter the variance components in a quadratic form. The squared sampling error may not converge to zero as the number of firms increases. Intuitively, the bias arises from an insufficient number of movers into and out of the firm, hence "limited mobility bias", so that variances are overstated and covariances understated (Andrews et al., 2008).

The magnitude of the bias is inversely related to the degree of connectivity of the firm-worker graph, with the graph being disconnected as limiting case (Jochmans and Weidner, 2019). For details, see Bonhomme et al. (2023, Section 3). Fortunately, the literature on panel data has made good progress in addressing this problem.

One approach defines the relevant level of firm unobserved heterogeneity as the "class" of a firm, corresponding to a cluster of similar employers (Bonhomme, Lamadon, and Manresa, 2019). While the class can be arbitrarily close to an individual firm, this may not be desirable because the number of job movers per firm will become smaller and result in an incidental parameters bias (i.e., reinstate the limited mobility problem). Under the assumptions of this approach, unobservable firm heterogeneity operates at the level of firm classes. The latter can be estimated in a first step through k-means clustering based on earnings and skills within each firm. This achieves two objectives: first, it enhances tractability; second, it delivers well-centered estimates of the contributions of worker and firm heterogeneity to earnings dispersion. Clustering trades

[3] Aghion et al. (2023) further show that cognitive military test scores similar to ours strongly predict whether an individual becomes an inventor in Finland, another important later in life outcome and closely related to our analyses in Section 6.2.

off restrictions on the dimensionality of the underlying groups for increased connectedness between firm classes. Notably, this method does not require shedding observations to generate a connected set.

A different approach builds on variance component estimators designed for unrestricted linear models with heteroscedasticity of unknown form. This removes the bias by resorting to leave-out estimators of the variances of errors from the linear model. For each worker–firm match, an estimate of the error variance is obtained from a sample where those observations are left out. The leave-out procedure delivers unbiased estimates in finite samples (Kline, Saggio, and Sølvsten, 2020) and facilitates tests of linear restrictions.

This estimation strategy requires that firms remain connected by worker mobility when any single mover is dropped, which involves pruning the original sample to ensure that the connectedness condition is met by all leave-out subsamples. This reduces dimensionality and computation time (see Appendix B).

Implementation of the clustering and bias correction approaches. Implementation of the grouping approach requires clustering firms into classes. We define 100 classes using a k-means algorithm based on average earnings as well as average cognitive and noncognitive skills of all workers (stayers and movers). Having a sufficiently large set of classes accommodates rich heterogeneity and ensures stability while still delivering a major dimension reduction. Using information beyond wages has been proposed in the structural literature (Eeckhout and Kircher, 2011; Hagedorn et al., 2017; Bartolucci et al., 2018; Bagger and Lentz, 2019). In our implementation this is further motivated by the theoretical restriction that firm-specific production arrangements affect both the skill composition of the workforce and their wages. Alternative clustering criteria (e.g., adding within-firm dispersion of skills and wages, or employment levels) as well as alternative numbers of classes deliver similar results (Section 6.3). The availability of skill measures makes it feasible to estimate specifications that feature firm effects in both levels and returns. Previous work has shown that class membership and fixed effects can be accurately estimated with sufficiently many workers. Using skill proxies also avoids incidental parameter biases in estimated returns due to few panel observations per worker.

Implementation of the bias correction approach relies on the leave-one-out double-connected set of firms. We prune the sample to contain firms that remain connected along both skill dimensions (cognitive and noncognitive) for different levels (high and low) when each single observation is dropped. The implementation accounts for correlation of error terms within an individual's spell at a given employer (Kline et al., 2020). This is done by averaging the data to the worker–firm match level. The resulting leave-match-out set is double-connected (in both skill dimensions) and smaller than the original sample but allows for estimation at the individual firm level. The extensive size of the Swedish population data assuages concerns about sample sizes. Appendix B discusses theoretical details of each approach and their implementation. In Table A.1 we report statistics for the underlying samples.

A.3 Estimation Samples

Clustered estimation: sample and firm grouping. We concentrate on the largest set of firms connected via worker mobility. This corresponds to moving from column (1) to (2) in Table A.1, and is in fact not strictly necessary: for estimating clusters only mobility between firm classes is required, a condition almost trivially satisfied here. Nonetheless, we keep with existing literature and require connected firms; this is not a consequential sample restriction, as shown in Table A.1. The latter finding indicates that even our initial restrictions are enough to lead to a sample of relatively large and well-connected firms. Overall, there are 25,783 unique firms and 510,077 workers who move between firms at least once during 1999–2008 in column (2) of the table.

Next, we employ the k-means algorithm (see also Bonhomme et al., 2019, Section 4) to group firms into 100 clusters. We do this by using variation in mean earnings, mean cognitive, and mean noncognitive skills, which reflect the dimensions of firm heterogeneity that we are interested in. In particular, differing technologies should lead to variation in both firms' skill composition and earnings. We estimate model (6) using this sample and the definition of firm clusters (i.e., the j subscripts refer to the 100 clusters). Results are reported in Column (1) of Table 1. Section 6.3 in the paper and associated Appendix F.3 examine robustness with respect to alternative clustering criteria as well as to the number of firm classes.

Bias-correction estimation: leave-one-out match-level samples. The estimation of variance components with the bias correction requires a set of firms that are leave-one-out connected by mobility of high and low skill workers in both the cognitive and noncognitive dimension. We meet this condition by only sampling firms that are leave-one-out connected through: (i) low skill workers ($c \leq 5, n \leq 5$), (ii) low in one and high in the other dimension workers ($c \geq 6, n \leq 5$ or $c \leq 5, n \geq 6$), and (iii) generally high-skill workers ($c \geq 6, n \geq 6$). A leave-one-out connected set of firms remains connected when any one worker is removed. This requires finding the workers that constitute cut vertices or articulation points in the corresponding bipartite network (Kline et al., 2020, Computational Appendix 2.1).

The algorithm to construct our estimation sample works as follows:

Step 1: We use Python's NetworkX package to identify the articulation points of the worker–firm graph, remove them and find the largest connected set that remains, then add back those articulation points that are connected to this largest leave-one-out connected set.

Step 2: We identify the largest leave-one-out connected set separately for the three skill groups (i)–(iii) and only keep those firms that are in the intersection of these sets.

We repeat Steps 1 and 2 until there is no reduction in the size of the graph (i.e., the three largest leave-one-out connected sets coincide). This final set is leave-one-out connected for the three skill groups.

We estimate the model at the worker–firm match level to account for potential serial correlation within worker–firm employment spells. That is, we collapse the data to means and drop workers who stay in the same firm throughout the period, since in the match-level estimation

Table A.1: Summary statistics for the estimation samples

	Full sample (1)	Largest connected (2)	Leave-one-out (3)	Match-level (4)
Number of observations	6,610,567	6,609,865	3,267,381	1,188,618
Number of stayers	578,146	578,146	-	-
Number of movers	510,077	510,077	477,424	477,424
Number of firms	25,839	25,783	19,085	19,085
Average log annual earnings	7.84	7.84	7.83	7.83
StDev log annual earnings	0.60	0.60	0.64	0.71
Average cognitive skill	5.28	5.28	5.44	5.44
Average noncognitive skill	5.13	5.13	5.23	5.23
Average age	37.32	37.32	36.35	36.35

Notes: Summary statistics for successively more restricted samples. Column (1) are all males aged 20–60 with non-missing employer, earnings, and test scores 1999–2008 at firms that exist at least five years with at least ten sample workers on average. Column (2) extracts the largest connected set of firms and their employees. Column (3) extracts the leave-one-out connected set of firms and removes workers who stay in the same firm in all years they are observed. Column (4) collapses the column (3) sample to worker–firm matches (summary statistics weighted by underlying frequencies). Earnings are real annual labor income in 2008 Swedish kronor. Cognitive and noncognitive scores are in Stanine scale. Our estimation samples are in bold font, (2) for clustering and (4) for the bias-correction approach.

these do not contribute to identifying the firm effects. We thereby follow exactly Kline et al. (2020, Appendix A)'s recommendations for estimating variance components in panels of $T > 2$.

The final firm-level sample to estimate (6) is summarized in Column (4) of Table A.1. This consists of 19,084 unique firms and 477,423 mover workers within the firm-level sample. The leave-one-out connectedness requirement increases employer size as it reduces the number of firms (26%) relatively more than the number of workers (7%). However, these reductions seem to have moderate effects. The average and dispersion of earnings do not change much but workers in the firm-level sample with larger firms are slightly younger and more skilled. The smaller number of observations in the match-level sample, without stayers and collapsed to the worker–firm match level, also reduces the computational burden (see footnote 6 below). For comparison, we also show results for the leave-one-observation-out sample in Table E.1 and, as expected, estimated dispersions of firm returns are substantially larger. In that sense, the match-level results in the main text are conservative.

B Overview of Econometric Methods

Throughout the paper we use high-dimensional firm effects specifications featuring firm-specific returns to cognitive and noncognitive skills. Estimates from these models are employed to study quadratic forms of model parameters. The baseline linear model is[4]

$$\log(w_{ijt}) = \mu_i + \lambda_j^0 + c_i \cdot \lambda_j^c + n_i \cdot \lambda_j^n + \varepsilon_{ijt}.$$

Of particular economic interest is the set of second moments of firm and worker specific parameters. For instance, in the standard double fixed effect model, one might interpret $\text{cov}(\mu, \lambda^0)$ as a measure of sorting of high-type workers into high-type firms. However, the naive plug-in estimates of these moments are prone to biases. In fact, developing unbiased estimators of such quadratic forms is the object of several papers in the firm heterogeneity literature (Andrews et al., 2008; Bonhomme et al., 2019; Kline et al., 2020; Bonhomme et al., 2023). Since our interest is in studying similar second moments, in what follows we briefly overview some details about the methods we employ to estimate firm effects.[5]

B.1 Estimating Bias-Corrected Quadratic Forms

We begin by rewriting our baseline specification as:

$$\begin{aligned} \log(w_{ijt}) &= \mu_i + \lambda_j^0 + c_i \cdot \lambda_j^c + n_i \cdot \lambda_j^n + \varepsilon_{ijt}, \\ &\equiv \mathbb{X}_{ij}\beta + \varepsilon_{ijt} \end{aligned} \tag{B.1}$$

where $\beta = [\mu; \lambda^0; \lambda^c; \lambda^n]' \equiv [\mu_1, ..., \mu_I;\ \lambda_1^0, ..., \lambda_J^0;\ \lambda_1^c, ..., \lambda_J^c;\ \lambda_1^n, ..., \lambda_J^n]'$ is the parameter vector and $\mathbb{X}_{ij} = [\mathbf{1}_i,\ \mathbf{1}_j,\ c_i\mathbf{1}_j,\ n_i\mathbf{1}_j]$ is the data matrix.

The symbol $\mathbf{1}_i$ denotes a $I \times 1$ indicator vector whose elements are all zero except the i^{th} coordinate (corresponding to worker i) which is set to 1. Similarly $\mathbf{1}_j$ is a $J \times 1$ indicator vector for firm j.

Kline et al. (2020) suggest an unbiased estimator for arbitrary quadratic forms involving the coefficients of (B.1) in the form of $\beta' A \beta$, for given matrix A. By appropriately choosing the A matrix, one can recast all the second moments of firm parameters λ_j^0, λ_j^c, and λ_j^n into quadratic expressions of the form $\beta' A \beta$.

Constructing quadratic forms. We begin by defining three row vectors associated to different firm parameters: $\mathbb{X}_{ij}^0 = [0_{1\times I},\ \mathbf{1}_j,\ 0_{1\times J},\ 0_{1\times J}]$, $\mathbb{X}_{ij}^c = [0_{1\times I},\ 0_{1\times J},\ \mathbf{1}_j,\ 0_{1\times J}]$, and $\mathbb{X}_{ij}^n = [0_{1\times I},\ 0_{1\times J},\ 0_{1\times J},\ \mathbf{1}_j]$, where i identifies worker and j is firm. Also, we let $\mathbb{X}$ denote the matrix that results from vertically stacking all the observations in row vector $\mathbb{X}_{ij}$. Then, $\mathbb{X}^0$, $\mathbb{X}^c$, and

[4] In the specifications studied in the main body we also include a broad set of control variables which are ignored here for notational simplicity.

[5] For in-depth discussions of these estimators see Kline et al. (2020) and Bonhomme et al. (2022, 2019).

$\mathbb{X}^n$ denote the matrices that result from vertically stacking $\mathbb{X}^0_{ij}$, $\mathbb{X}^c_{ij}$ and $\mathbb{X}^n_{ij}$. Finally, we define

$$A^0 = \frac{1}{\sqrt{N}} \left(\mathbb{X}^0 - \overline{\mathbb{X}}^0 \right)$$
$$A^c = \frac{1}{\sqrt{N}} \left(\mathbb{X}^c - \overline{\mathbb{X}}^c \right)$$
$$A^n = \frac{1}{\sqrt{N}} \left(\mathbb{X}^n - \overline{\mathbb{X}}^n \right)$$

where $\overline{\mathbb{X}}^0 = \frac{1}{N}[0_{N\times I},\ 1_{N\times J},\ 0_{N\times J},\ 0_{N\times J}]$, $\overline{\mathbb{X}}^c = \frac{1}{N}[0_{N\times I},\ 0_{N\times J},\ 1_{N\times J},\ 0_{N\times J}]$, and $\overline{\mathbb{X}}^n = \frac{1}{N}[0_{N\times I},\ 0_{N\times J},\ 0_{N\times J},\ 1_{N\times J}]$. One can use the A matrices above to estimate second moments of interest, e.g. $\text{VAR}(\lambda^0) = \beta'(A^{0\prime}A^0)\beta$ or $\text{COV}(\lambda^c,\lambda^n) = \beta'(A^{c\prime}A^n)\beta$. In what follows we set $A = A_1'A_2$ to estimate $\theta = \beta'A\beta$, where A_1 and A_2 could be any of A^0, A^c, and A^n (depending on which moments we are interested in).

Plug-in estimator. The plug-in estimator $\hat{\theta}_{\text{PI}} = \hat{\beta}'A\hat{\beta}$ can be obtained by simply using the OLS estimates of $\hat{\beta}$ in the quadratic form defining θ. However, the plug-in estimator is biased and its expected value is

$$\mathbb{E}[\hat{\theta}_{\text{PI}}] = \theta + \text{trace}(A \times \text{VAR}[\hat{\beta}]) = \theta + \sum_{k=1}^{N} B_{kk}\sigma_k^2 \tag{B.2}$$

where $S = \mathbb{X}'\mathbb{X}$, B_{kk} is the k-th diagonal element of $B = \mathbb{X}S^{-1}AS^{-1}\mathbb{X}'$ corresponding to observation k, and σ_k^2 is the variance of error term of observation k. Therefore, the bias in the plug-in estimator can be corrected by using unbiased estimates of σ_k^2, which is the route we take when estimating the model at the level of individual firms.

Bias-corrected quadratic forms. We use leave-k-out OLS estimators of β, denoted by $\hat{\beta}_{-k}$, that are obtained from a sample where the observation k is excluded. This delivers an unbiased estimator of σ_k^2 such that

$$\hat{\sigma}_k^2 = y_k(y_k - x_k\hat{\beta}_{-k}), \tag{B.3}$$

where y_k is the dependent variable (i.e. log earnings) of observation k and x_k is the corresponding independent variables vector (i.e. row k of $\mathbb{X}$). Using the $\hat{\sigma}_k^2$ above, we compute the bias corrected estimator of θ as

$$\hat{\theta}_{KSS} = \hat{\beta}'A\hat{\beta} - \sum_{k=1}^{N} B_{kk}\hat{\sigma}_k^2. \tag{B.4}$$

Large Scale Computations. Estimating $\hat{\theta}_{KSS}$ is computationally expensive for large data-sets with many estimated parameters such as ours. Like Kline et al. (2020), we use a variant of the random projection method of Achlioptas (2003, known as Johnson-Lindenstrauss Approximation, or JLA) to estimate the $\hat{\sigma}_k^2$ and B_{kk} required in the estimation of $\hat{\theta}_{KSS}$. JLA suggests the following approximation:

$$\hat{P}_{kk} = \frac{1}{p}||R_P \mathbb{X} S^{-1} x_k||^2$$

$$\hat{B}_{kk} = \frac{1}{p}(R_B A_1 S^{-1} x_k)'(R_B A_2 S^{-1} x_k)$$

$$\hat{\sigma}^2_{k,JLA} = \frac{y_k(y_k - x_k\hat{\beta})}{1-\hat{P}_{kk}}(1 - \frac{1}{p}\frac{3\hat{P}^3_{kk} + \hat{P}^2_{kk}}{1-\hat{P}_{kk}}),$$

where $p \in \mathbb{N}$ is a number much smaller than the total number of estimated parameters. That is, we can achieve a material reduction in the dimensionality of the problem. The $R_P, R_B \in \{-1,1\}^{p\times N}$ are random matrices of order $p \times N$ featuring elements equal to +1 and -1 with equal probabilities. This makes computations significantly faster when parameters are estimated at the level of individual firms.[6]

B.2 Cluster-Based Estimation

Models with two sided heterogeneity rely on job movers to identify the unobserved firm and worker parameters. In typical employer–employee linked data sets the number of job movers per firm tends to be small, which leads to the well known limited mobility bias in quadratic forms of these estimates. To alleviate this problem, group based estimates have been suggested in the literature. In this approach, firm parameters are assumed to only vary across groups or clusters of firms, rather than individual firms. Under this assumption about the underlying data generating process and further assuming that the number of groups is limited, the number of job moves per group of firms is sufficiently large, which alleviates the small sample bias concern.

Partitioning returns across clusters. To adapt this framework to our setting, we begin by rewriting the baseline specification as

$$\log(w_{ijt}) = \mu_i + \lambda^0_{g(j)} + c_i \cdot \lambda^c_{g(j)} + n_i \cdot \lambda^n_{g(j)} + \varepsilon_{ijt},$$

where $g : \{1,...,J\} \to \{1,...,K\}$ is a partitioning function that maps firm j into cluster $g(j)$ that the firm j belongs to, and K is the total number of groups. These groups could in principle be the individual firms, i.e., $g(j) = j$, but only in models with a reduced number of groups is the limited mobility bias less of a concern.

[6] Estimating the bias-corrected second moments of parameters in model (6) on the data in Column (4) of Table A.1 takes about 20–30 hours using Python and the JLA approximation with $p = 50$ depending on the Swedish server's workload. Setting $p = 50$ is in line with Kline et al. (2020) and we have tested that further increasing p does not change our results.

Two-step estimation. We estimate the model in two steps i.e. (i) partition firms into K disjoint groups (ii) estimate the model using the firm groups. Bonhomme and Manresa (2015) show that a k-means estimator can consistently identify the firm classes up to a relabeling of groups. In the first step, as discussed in Section 2.2, we use the average earnings as well as average cognitive and noncognitive traits of their workers to group firms. Intuitively, the earnings and average skills in firms with identical intercepts and returns should be the same and one could then use these observed firm variables to define separate firm classes. The structural literature advocates going beyond earnings when clustering firms (Eeckhout and Kircher, 2011; Hagedorn et al., 2017; Bartolucci et al., 2018; Bagger and Lentz, 2019), since a classification may fail to be identified when two firm classes have identical earnings distributions in the cross section.[7]

Once firm groups are defined, firm and worker parameters are identified (up to the normalization discussed in Section 3.1 of the main text) under the assumptions of serial conditional independence of earnings and random job mobility (Bonhomme et al., 2019), and estimated using panel regressions in conjunction with skill proxies.

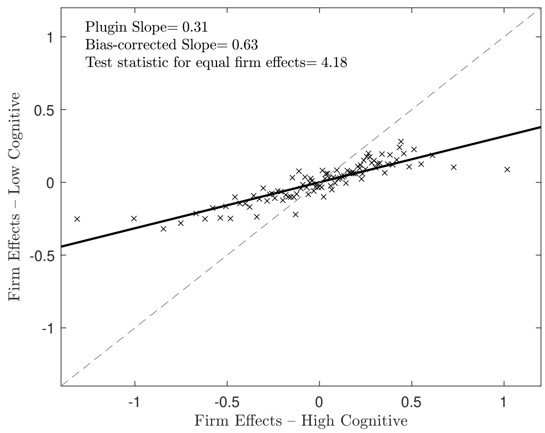
B.3 Basic Test of Returns Heterogeneity: Bias-Corrected Slopes

In Section 2.3 of the paper we test the hypothesis that firm effects are independent of worker skills using an additive specification with binary skill levels (high vs low test scores) for each skill attribute. To run these tests we construct subsamples corresponding to the largest connected sets of, respectively, high and low ability workers and select firms that are in both of these sets (double-connectedness in skill levels).

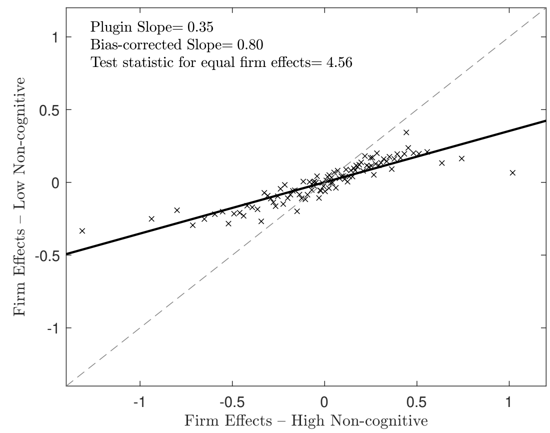
We first classify workers into high cognitive (Stanine $C = \mathbb{1}[c > 5]$) and high noncognitive ($N = \mathbb{1}[n > 5]$); then, we select observations within a two-year set to separately estimate linear binary models of worker and firm effects of the form $\log(w_{ijt}) = \mu_i^S + \theta_j^S + \varepsilon_{ijt}$ for cognitive skills, $S \in \{C = 0, C = 1\}$ or noncognitive skills, $S \in \{N = 0, N = 1\}$. Figure 1 in the main body plots shows results when grouping firms into 100 clusters for $t \in \{2004, 2007\}$ (see Bonhomme et al., 2019). We use non-adjacent years (in fact, two years apart) to mitigate the impact of partial employment spells during contiguous years when workers switch firms. Independence of skill premia from skills would not be rejected if the estimated slopes in the scatter plots were not significantly different from one in a standard t-test. However, the null hypothesis that firm effects are the same for high- and low-skill workers are strongly rejected for both cognitive and noncognitive skills.

The clustering accounts for the incidental parameter bias due to limited worker mobility. To assess the robustness of these findings we run the same tests using bias-corrected slopes (see Kline et al., 2020). Since the analysis is carried out separately for cognitive and noncognitive attributes, double-connectedness in skill levels does not (yet) require that firms be linked through mobility of both skill dimensions (in other sections we examine set connectedness for the case where multiple skills are considered in the same specification).

[7] For example, a firm class may have higher intercepts and the other higher returns but worker sorting is such that observed earnings are the same. See also discussion in Bonhomme et al. (2019, page 14).



(A) High vs low cog skills – leave-one-out correction



(B) High vs low noncog skills – leave-one-out correction

Figure B.1: Firm effects heterogeneity: cognitive and noncognitive skills. Bias-Corrected Slopes.
These figures plot the averages of firm effects for low-skill workers ($\theta_j^{S=0}$) against the averages of firm effects for high skill workers ($\theta_j^{S=1}$), where $S \in \{C,N\}$. All sets of firm effects are demeaned. The sample in panel (A) consists of 9,268 firms that are leave-one-out connected in both high and low cognitive skills; in panel (B) we use 10,208 firms connected in both high and low noncognitive skills. The "plugin slope" is the coefficient from a person-year weighted projection of $\theta_j^{S=0}$ onto $\theta_j^{S=1}$. The "bias-corrected slope" adjusts the plug-in slope for attenuation bias by multiplying its value by the ratio of the plug-in estimate of the person-year weighted variance of $\theta_j^{S=1}$ to the bias-adjusted estimate of the same quantity. "Test Statistic" refers to the realization of $\hat{z}_{H_0}/\sqrt{v\hat{a}r(\hat{z}_{H_0})}$ where $\hat{z}_{H_0}$ is the quadratic form associated with the null hypothesis that the firm effects are equal across skill groups. From Theorem 2 in Kline et al. (2020), $\hat{z}_{H_0}/\sqrt{v\hat{a}r(\hat{z}_{H_0})}$ converges to a $N \sim (0,1)$ under the null hypothesis that $\theta_j^{S=0} = \theta_j^{S=1}$ for, respectively, all 9,268 and 10,208 firms.
Sample restriction: years 2004 and 2007 only. Tests for other year pairs are in Table B.1.

Panels (A)–(B) in Figure B.1 plot a scatter of estimated firm fixed effects for high-skill (x-axis) and low-skill (y-axis) workers. The samples consist of firms that are in the leave-one-out connected sets of both high and low ability workers. Each panel refers to a given skill attribute, covering the years 2004 and 2007. Panel (A) shows results for cognitive skills (9,268 firms) while Panel (B) plots those for noncognitives (10,208 firms).

A comparison of firm effects (θ_j^S) for high and low skill workers illustrates that the statistics of firm effects (like their variances and correlations) can be biased if estimated from few moves of workers into and out of each firm. Ignoring estimation biases results in firm effects for high and low skill workers that are positively but weakly correlated within firms. The regression slope from mechanically projecting $\theta_j^{S=0}$ onto $\theta_j^{S=1}$ is 0.31 for cognitive traits and 0.35 for noncognitives. We refer to these slopes as the "plug-in" estimates. The "bias-corrected slope" adjusts the plug-in slope for attenuation bias by multiplying its value with the ratio of the plug-in estimate of the person-year weighted variance of $\theta_j^{S=1}$ to the bias-adjusted estimate of the same quantity. The "Test Statistic" is the realization of $\hat{z}_{H_0}/\sqrt{v\hat{a}r(\hat{z}_{H_0})}$ where $\hat{z}_{H_0}$ is the quadratic form associated with the null hypothesis that the firm effects are equal across skill groups. From Theorem 2 in Kline et al. (2020), $\hat{z}_{H_0}/\sqrt{\hat{var}(\hat{z}_{H_0})}$ converges to a $N \sim (0,1)$ under the null hypothesis that $\theta_j^{S=0} = \theta_j^{S=1}$ for, respectively, all 9,268 and 10,208 firms.

The bias correction raises estimated slopes to 0.63 and 0.81, respectively. Under the null hypothesis of no heterogeneity in skill returns, however, the slopes should be statistically indistinguishable from one and the scatters should align along the dashed 45° lines. This is not the case, as the bias-corrected test statistics of equal firm effects for high and low skill workers have z-values above 4 for both cognitive and noncognitive returns. We therefore reject the hypothesis that firm effects are independent of worker skills. In fact, all our estimates indicate slopes that are well below one. Table B.1 reports additional tests, which similarly reject the null hypothesis of homogeneous returns in several alternative samples.[8]

Table B.1: Tests for equality of firm effects by high- versus low-skill workers (by year combination and cognitive / noncognitive)

Year origin (1)	Year destination (2)	Skill (3)	Test Statistic Firm-level (4)	# Firms (5)	Test Statistic Grouped (6)
1999	2002	C	3.66	8,757	9.22
1999	2002	N	2.49	9,766	12.45
2000	2003	C	2.60	8,653	8.66
2000	2003	N	0.39	9,648	8.14
2001	2004	C	2.76	7,922	9.93
2001	2004	N	1.78	8,941	7.65
2002	2005	C	0.60	7,904	10.83
2002	2005	N	3.50	8,772	6.96
2003	2006	C	4.04	8,335	13.88
2003	2006	N	0.85	9,258	7.00
2004	2007	C	4.18	9,269	17.33
2004	2007	N	4.56	10,209	6.30
2005	2008	C	3.26	9,846	10.74
2005	2008	N	2.54	10,825	5.38

Notes: Table B.1 expands on Figure 1 to show test statistics associated with the null hypothesis that firm effects ($\theta_j^{S=0}$) and ($\theta_j^{S=1}$) are equal across skill level, where skill $S \in \{C, N\}$. Test statistic for firm-level bias-adjusted estimates as in Figure 1(a)–(b) are shown in column (4). The associated number of double-connected firms in each of the skill types and year combinations are reported in column (5). The last column reports the corresponding test statistic among 100 firm classes using the clustering approach as in Figure 1(c)–(d).

[8] Tests of equal slopes are based on an upper bound for the estimated error variance $\text{var}(\varepsilon_{ijt})$. This leads to conservative test statistics compared to the split-sample estimate in Figure 1 of Kline et al. (2020). Joint tests of the equal effects hypothesis across more than two periods are unfeasible as they introduce issues with clustering of errors at the firm level. No robust procedure is currently available to handle such issues. We thank Raffaele Saggio for discussions about implementing these tests.

Testing equality of firm effects in alternative year pairs. In Section 2.3, and Figure B.1, we choose the years 2004 and 2007 to test the equality of firm effects for high versus low skilled workers. Two years are selected to exclude potential serial correlation within employment spells due to estimated standard errors (see Kline et al., 2020, Computational Appendix 2.5).

Years are non-adjacent, in order to remove partial employment years when workers switch firms, while not too far apart to minimize any potential changes in firm effects over time. The sample is selected on firms that are leave-one-out connected in both high and low levels of the respective skill dimension ("double-connected").

To gauge robustness, we also replicate the analysis for alternative duplets of years. Like in Section 2.3, we focus on testing the null hypothesis that firm effects are equal for high and low skills groups; the hypotheses are separately tested for cognitive and noncognitive skills. Table B.1 shows the resulting test statistics and sample sizes of the respective double-connected individual firms for several year pairs using the bias-correction approach. The last column reports the corresponding test statistics among 100 firm classes using clustering as in the main text.

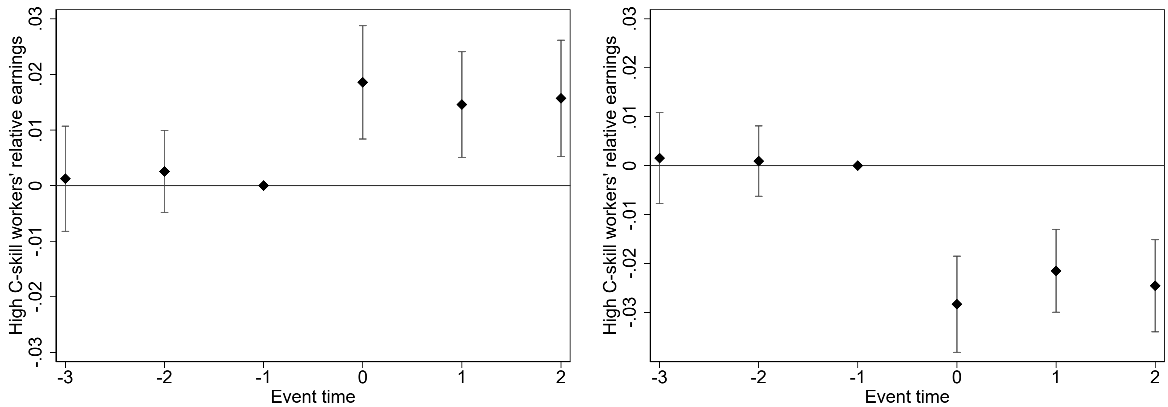
B.4 Moving Across Firms: Event Studies

Heterogeneity in skill returns can be further illustrated through event studies that track wage changes for different workers as they move across firms (e.g., Bonhomme et al., 2019; Lamadon et al., 2022). In what follows, we estimate wage changes for workers of different skills who make similar switches.

We consider a balanced sample of workers observed for at least three years in the origin firm before their move and require that workers be observed for at least three years in the destination firm after the switch. Next, for each attribute we define high-skill dummies $C = \mathbb{1}[\tilde{c} > 0]$ and $N = \mathbb{1}[\tilde{n} > 0]$, where $\tilde{c}$ and $\tilde{n}$ are mean zero residualized cognitive and noncognitive skills. [9] To allay endogeneity concerns, we construct firm-level skill premia using coworkers' wages. That is, in each year we compute the difference in the average log earnings between high-skill and low-skill coworkers within each firm (done separately for cognitive and noncognitive attributes). Then, we average this difference over all the firm's years.

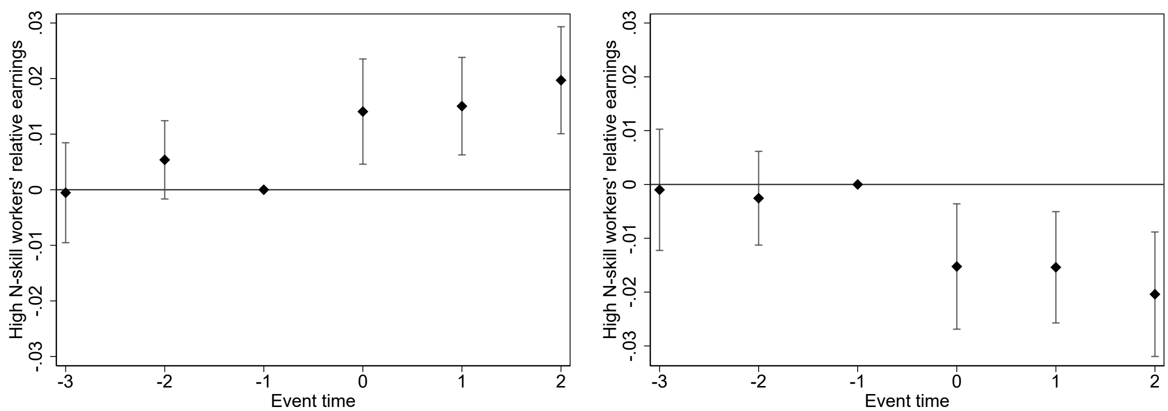
Panel (A) of Figure B.2 shows the effect of workers moving to a new firm where the coworker cognitive skill premium is at least ten log points higher than in the origin firm. The variable of interest is the wage of high-skill workers relative to low-skill ones making the move. No pre-trend is visible in relative earnings; we do find a significant and discontinuous increase upon entry into the higher-return firm and, thereafter, a flattening of relative earnings around a level that is 1.5–2 log points above their pre-switch value. These estimates lend support to the hypothesis of significant differences in skill returns. Estimates are almost a mirror image when we examine workers switches to a firm with lower cognitive skill premia (namely, at least ten log points below the origin firm), shown in Panel (B) of Figure B.2. The point estimate of the difference is close to three log points. That is, skilled workers' relative wages decline when they move to lower-return

[9] In the baseline event studies we residualize each skill measure (cognitive or noncognitive) with respect to the other in order to reduce possible confounding effects due to covariation between attributes.



(A) Switch to firm with higher coworker *C*-premium

(B) Switch to firm with lower coworker *C*-premium



(C) Switch to firm with higher coworker *N*-premium

(D) Switch to firm with lower coworker *N*-premium

Figure B.2: Event studies of relative wage changes for high versus low skill workers after switching across firms with different coworker skill premia.

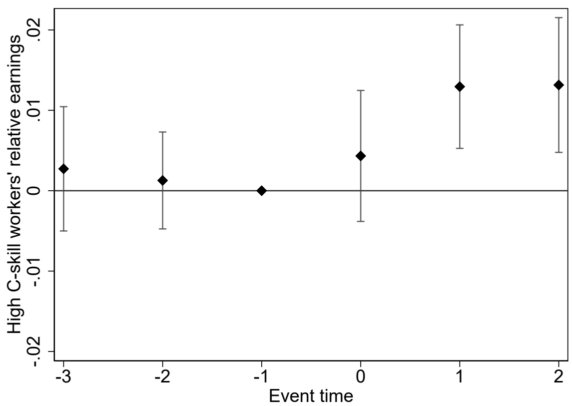
Notes: Top panels: cognitive skills (C). Bottom panels: noncognitive skills (N). Switches occur between $t = -1$ (last year in old firm) and $t = 0$ (first year in new firm). Balanced sample of moves with standard errors clustered at the worker level. 95% confidence intervals drawn around point estimates.

firms. Wage changes after switches are qualitatively similar along the noncognitive dimension, shown in the bottom panels of Figure B.2. While magnitudes are smaller, relative earnings are impacted also in the noncognitive dimension.

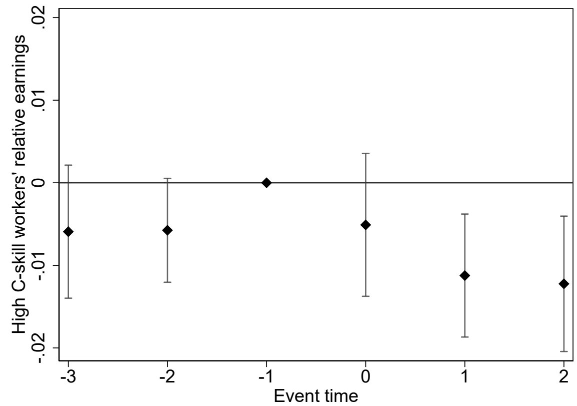
Overall, the event studies indicate roughly symmetric relative wage effects and no clear pre- or post-trends. Estimates are comparable in magnitude to those in Fredriksson et al. (2018) who use coworkers' skills to examine the effects of job mismatch at one year lags and beyond. In the following we explore alternative implementations of the event studies where we define skill premia using model estimates of heterogeneous returns taken from Section 3 (as opposed to coworker wages). Findings based from these exercises indicate patterns of firm-level heterogeneity that are close, both qualitatively and quantitatively, to estimates from the full model of Section 3. Moreover, we report evidence that skill returns are attribute-specific, in the sense that relative wage gains for cross-moves (e.g., high-versus-low cognitive workers switching to higher noncognitive premium firms) are close to zero and not significant.

Event studies: Additional evidence and extensions. Figure B.3 shows estimates of relative wage changes in event studies of moves between higher and lower skill returns λ_j^c, λ_j^n firms, where skill premia are defined according to returns estimated in the full model (6). These estimates are qualitatively similar to those shown above and provide a more direct interpretation of the magnitudes of estimated effects: the moves between high and low returns firms are associated with changes of $\Delta\lambda_j^c \approx 0.05$ and $\Delta\lambda_j^n \approx 0.04$ on average in the respective dimension (not visible in the figure). On average the $\tilde{c}$ and $\tilde{n}$ residual skill differences between high- and low-skill classified workers are about 0.3. Therefore, the average effect sizes implied by the full model (6) for such switches are approximately $0.05 \cdot 0.3 = 0.015$ and $0.04 \cdot 0.3 = 0.012$, respectively. Estimates of changes in relative wages from the event studies in Figure B.3 are almost as large as model-based estimates: about 0.012–0.013 in the top panels for the cognitive skill dimension and 0.008–0.010 in the bottom panels for noncognitives. Also these event studies exhibit no clear pre-trends before the switch and broadly symmetric relative wage effects after the moves to higher-return firms vis-a-vis lower-return firms.

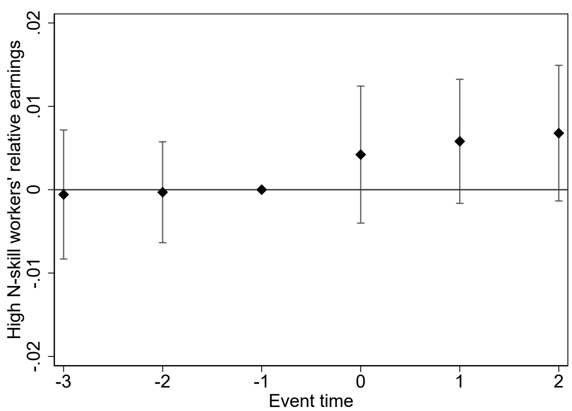
Finally, Figure B.4 illustrates estimated cross-effects from event studies based on the coworker wage measure of skill premia (that is, the same measures and approach as in Figure B.2). Panel (A) displays the wage change of a high (residualized) *cognitive* skill $\tilde{c}$ worker, relative to a low skill one, moving to a new firm where the coworker *noncognitive* skill premium is at least ten log points higher than in the old firm. The estimated wage change is small and insignificant. Similarly, in Panel (B) moving to a lower noncognitive premium firm has essentially zero effect in the initial year ($t = 0$) and exhibits negative and insignificant point estimates of wage impacts in the years $t = 1$ and $t = 2$. Relative wage effects for high (residualized) noncognitive skill $\tilde{n}$ individuals moving to firms with higher (Panel C) or lower (Panel D) cognitive coworker skill premia are again flat and close to zero. Compared with the significant positive own-effects of Figure B.2, this is evidence of skill return heterogeneity across firms that is mostly attribute-specific rather than indirect.



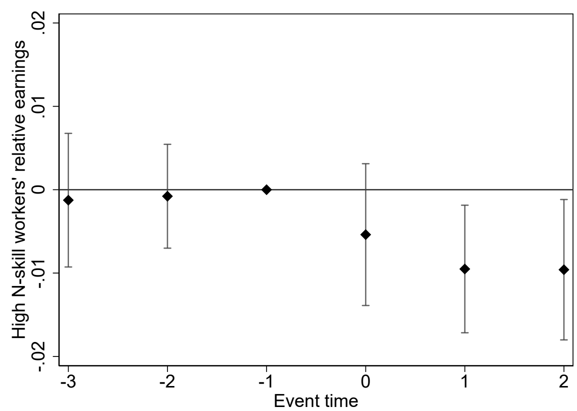
(A) Switch to firm with higher coworker *C*-premium



(B) Switch to firm with lower coworker *C*-premium



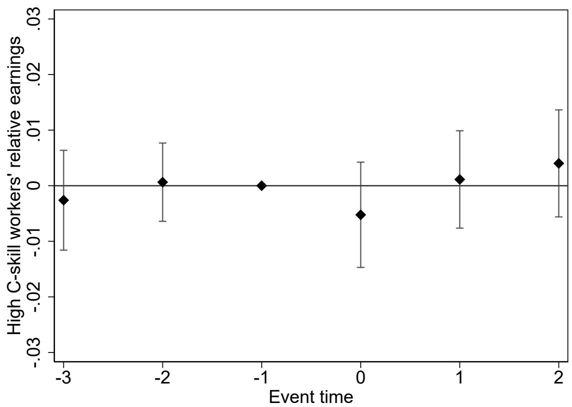
(C) Switch to firm with higher coworker *N*-premium



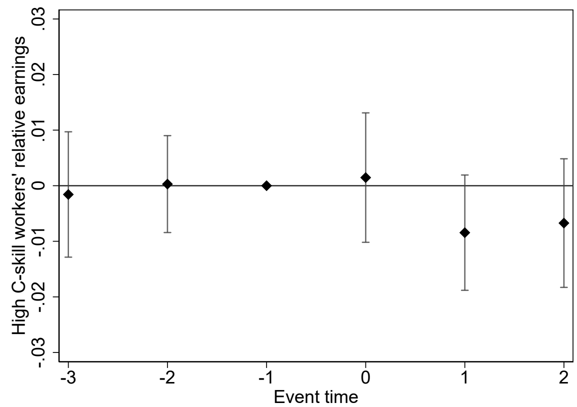
(D) Switch to firm with lower coworker *N*-premium

Figure B.3: Event studies of relative wage changes for high versus low skill workers after switching across firms with different estimated skill returns (λ_j^c and λ_j^n).

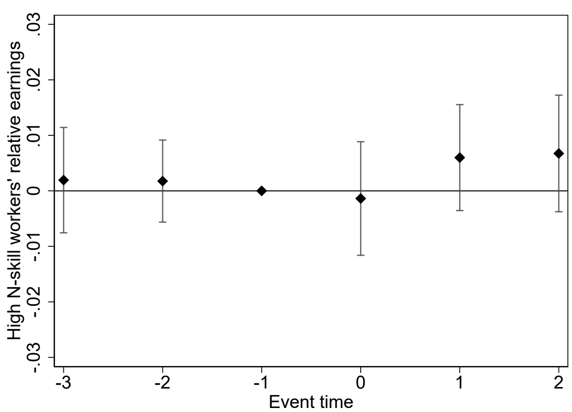
Notes: Top panels: cognitive skills (C). Bottom panels: noncognitive skills (N). Switches occur between $t = -1$ (last year in old firm) and $t = 0$ (first year in new firm). Balanced sample of moves with standard errors clustered at the worker level. 95% confidence intervals drawn around point estimates.



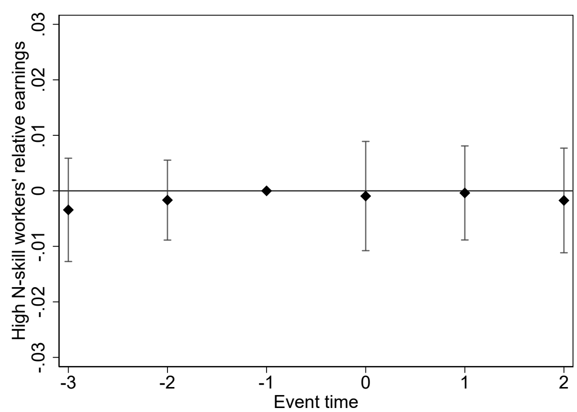
(A) Switch to firm with higher coworker *C*-premium



(B) Switch to firm with lower coworker *C*-premium



(C) Switch to firm with higher coworker *N*-premium



(D) Switch to firm with lower coworker *N*-premium

Figure B.4: Event study of cross effects: relative wage changes for high versus low skill workers after switching to a higher skill premium firm (in the respective other skill dimension).

Notes: Top panels: cognitive skills (C). Bottom panels: noncognitive skills (N). Switches occur between $t = -1$ (last year in old firm) and $t = 0$ (first year in new firm). Balanced sample of moves with standard errors clustered at the worker level. 95% confidence intervals drawn around point estimates.

B.5 Identification of Complementarities with Unobserved Skills

Complementarities between productivity attributes and employers are not generally identified if skills have more than one dimension unless an observable skill proxy is available. For illustration, consider a wage equation featuring firm-level returns to cognitive (c) and noncognitive (n) traits:

$$\log(w_{it}) = \lambda_j^0 + \lambda_j^{\mathrm{c}} c_i + \lambda_j^{\mathrm{n}} n_i + \varepsilon_{it} \tag{B.5}$$

Throughout this section we maintain the assumptions of connectedness (that is, we assume that employers are connected through workers' flows). We begin with the case of unobserved skills and examine identification in a large sample consisting of workers who move from firm j to j' and from j' to j between two periods $t = 1,2$ (Bonhomme et al., 2019, BLM). For clarity, we restrict attention to the simplest case where only two firms j and j' are present and show that returns are not generally identified if there are two skill dimensions.

Identification of heterogeneous returns in the absence of skill measures. Let $\mathrm{E}_{jj'}[.]$ denote the expectation of the argument over the workers who move from firm j to j' between $t = 1$ and $t = 2$. For instance, $\mathrm{E}_{jj'}[\log(w_{i1})]$ denotes the average log wage of workers who moved from j to j' at the end of period 1 (i.e., when working at firm j). Lacking measures of c_i and n_i, only the following four average wages can be defined from worker moves across firms:

$$\begin{aligned}
\mathrm{E}_{jj'}[\log(w_{i1})] &= \lambda_j^0 + \lambda_j^{\mathrm{c}} \mathrm{E}_{jj'}[c_i] + \lambda_j^{\mathrm{n}} \mathrm{E}_{jj'}[n_i] \\
\mathrm{E}_{j'j}[\log(w_{i2})] &= \lambda_j^0 + \lambda_j^{\mathrm{c}} \mathrm{E}_{j'j}[c_i] + \lambda_j^{\mathrm{n}} \mathrm{E}_{j'j}[n_i] \\
\mathrm{E}_{jj'}[\log(w_{i2})] &= \lambda_{j'}^0 + \lambda_{j'}^{\mathrm{c}} \mathrm{E}_{jj'}[c_i] + \lambda_{j'}^{\mathrm{n}} \mathrm{E}_{jj'}[n_i] \\
\mathrm{E}_{j'j}[\log(w_{i1})] &= \lambda_{j'}^0 + \lambda_{j'}^{\mathrm{c}} \mathrm{E}_{j'j}[c_i] + \lambda_{j'}^{\mathrm{n}} \mathrm{E}_{j'j}[n_i]
\end{aligned}$$

Or in matrix notation:

$$\begin{bmatrix} 1 & \mathrm{E}_{jj'}[c_i] & 0 & 0 \\ 1 & \mathrm{E}_{j'j}[c_i] & 0 & 0 \\ 0 & 0 & 1 & \mathrm{E}_{jj'}[c_i] \\ 0 & 0 & 1 & \mathrm{E}_{j'j}[c_i] \end{bmatrix} \begin{bmatrix} \lambda_j^0 \\ \lambda_j^{\mathrm{c}} \\ \lambda_{j'}^0 \\ \lambda_{j'}^{\mathrm{c}} \end{bmatrix} + \begin{bmatrix} \mathrm{E}_{jj'}[n_i] & 0 \\ \mathrm{E}_{j'j}[n_i] & 0 \\ 0 & \mathrm{E}_{jj'}[n_i] \\ 0 & \mathrm{E}_{j'j}[n_i] \end{bmatrix} \begin{bmatrix} \lambda_j^{\mathrm{n}} \\ \lambda_{j'}^{\mathrm{n}} \end{bmatrix} = \begin{bmatrix} \mathrm{E}_{jj'}[\log(w_{i1})] \\ \mathrm{E}_{j'j}[\log(w_{i2})] \\ \mathrm{E}_{jj'}[\log(w_{i2})] \\ \mathrm{E}_{j'j}[\log(w_{i1})] \end{bmatrix} \tag{B.6}$$

Lemma 1. *Suppose that* $\mathrm{E}_{jj'}[c_i] \neq \mathrm{E}_{j'j}[c_i]$*; then we can write*

$$\begin{aligned}
\begin{bmatrix} \lambda_j^0 \\ \lambda_j^{\mathrm{c}} \\ \lambda_{j'}^0 \\ \lambda_{j'}^{\mathrm{c}} \end{bmatrix} = & \frac{1}{\mathrm{E}_{j'j}[c_i] - \mathrm{E}_{jj'}[c_i]} \begin{bmatrix} \mathrm{E}_{j'j}[c_i] & \mathrm{E}_{jj'}[c_i] & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & \mathrm{E}_{j'j}[c_i] & \mathrm{E}_{jj'}[c_i] \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mathrm{E}_{jj'}[\log(w_{i1})] \\ \mathrm{E}_{j'j}[\log(w_{i2})] \\ \mathrm{E}_{jj'}[\log(w_{i2})] \\ \mathrm{E}_{j'j}[\log(w_{i1})] \end{bmatrix} \\
& - \frac{1}{\mathrm{E}_{j'j}[c_i] - \mathrm{E}_{jj'}[c_i]} \begin{bmatrix} \left(\mathrm{E}_{j'j}[c_i]\mathrm{E}_{jj'}[n_i] - \mathrm{E}_{jj'}[c_i]\mathrm{E}_{j'j}[n_i]\right) \lambda_j^{\mathrm{n}} \\ \left(-\mathrm{E}_{jj'}[n_i] + \mathrm{E}_{j'j}[n_i]\right) \lambda_j^{\mathrm{n}} \\ \left(\mathrm{E}_{j'j}[c_i]\mathrm{E}_{jj'}[n_i] - \mathrm{E}_{jj'}[c_i]\mathrm{E}_{j'j}[n_i]\right) \lambda_{j'}^{\mathrm{n}} \\ \left(-\mathrm{E}_{jj'}[n_i] + \mathrm{E}_{j'j}[n_i]\right) \lambda_{j'}^{\mathrm{n}} \end{bmatrix}
\end{aligned}$$

Remark 1. *Suppose that* $\mathrm{E}_{jj'}[c_i] \neq \mathrm{E}_{j'j}[c_i]$. *Then, if* $\lambda^{\mathrm{n}}_j = \lambda^{\mathrm{n}}_{j'} = 0$ *or* $\mathrm{E}_{jj'}[n_i] = \mathrm{E}_{j'j}[n_i]$, *the ratio of returns across two firms* j *and* j' *can be written as*

$$\frac{\lambda^{\mathrm{c}}_{j'}}{\lambda^{\mathrm{c}}_j} = \frac{\mathrm{E}_{jj'}[\log(w_{i2})] - \mathrm{E}_{j'j}[\log(w_{i1})]}{\mathrm{E}_{jj'}[\log(w_{i1})] - \mathrm{E}_{j'j}[\log(w_{i2})]} \tag{B.7}$$

Equation (B.7) is constructed by picking the second and fourth row on the right-hand side of Lemma 1 and dividing them by each other. Unsurprisingly, this shows that, if $\lambda^{\mathrm{n}}_j = \lambda^{\mathrm{n}}_{j'} = 0$, the specification in (B.5) reduces to a wage equation with complementarities in the one-dimensional skill model (BLM, 2019). By the same token, if $\mathrm{E}_{jj'}[n_i] = \mathrm{E}_{j'j}[n_i] = \bar{n}$, there is no variation in the average noncognitive skill dimension of movers and the noncognitive return is absorbed into the intercept. In this case, with a change of variable $\tilde{\lambda}^0_j = \lambda^0_j - \bar{n}\lambda^{\mathrm{n}}_j$, the model again reduces to a one-dimensional skill model (single index representation).

Equation (B.7) is a restatement of an established identification result about the ratio of returns with a single unobserved skill dimension (BLM, 2019). The ratio of returns to unobserved skills is identified from differences in the average log wages of movers between firms j and j', even though the absolute returns are not generally identified. The condition $\mathrm{E}_{jj'}[c_i] \neq \mathrm{E}_{j'j}[c_i]$ requires that the average skill of movers from j to j' is not the same as that of movers from j' to j. This is necessary for identification in a model with a single index for unobserved skills.

Remark 2. *Suppose that* $\mathrm{E}_{jj'}[c_i] \neq \mathrm{E}_{j'j}[c_i]$; *then without further restrictions it follows that*

$$\frac{\lambda^{\mathrm{c}}_{j'}}{\lambda^{\mathrm{c}}_j} = \frac{\mathrm{E}_{jj'}[\log(w_{i2})] - \mathrm{E}_{j'j}[\log(w_{i1})] - \left(\mathrm{E}_{jj'}[n_i] - \mathrm{E}_{j'j}[n_i]\right)\lambda^{\mathrm{n}}_{j'}}{\mathrm{E}_{jj'}[\log(w_{i1})] - \mathrm{E}_{j'j}[\log(w_{i2})] - \left(\mathrm{E}_{jj'}[n_i] - \mathrm{E}_{j'j}[n_i]\right)\lambda^{\mathrm{n}}_j} \tag{B.8}$$

That is, the ratio of returns to cognitive skills are not generally identified without recourse to observable skill measures. The ratio of returns to noncognitive skills are not identified for the same reason. The intuition is that, with multiple skill dimensions, the ranking of workers is not unique; that is, a given difference in wages across firms may obtain from returns to different attributes. Put differently, a given estimate $\lambda^{\mathrm{c}}_{j'}/\lambda^{\mathrm{c}}_j$ based on (B.8) would depend on unknown unobserved skill differences ($\mathrm{E}_{jj'}[n_i] - \mathrm{E}_{j'j}[n_i]$) and returns ($\lambda^{\mathrm{n}}_j, \lambda^{\mathrm{n}}_{j'}$).

Identification of heterogeneous returns when skill measures are available. Equation (B.8) is identified when skill measures are available, as one can simply restrict attention to worker flows for which $\mathrm{E}_{jj'}[n_i] = \mathrm{E}_{j'j}[n_i]$. One can also ensure that the condition $\mathrm{E}_{jj'}[c_i] \neq \mathrm{E}_{j'j}[c_i]$ indeed holds for the same worker flows. In fact, the returns λ^{c}_j and $\lambda^{\mathrm{c}}_{j'}$ are separately identified by picking the respective row from the right-hand side of Lemma 1 and conditioning on the appropriate skill levels. In practice, we can identify a more general version of model (B.5) by relying on available skill measures; specifically, we consider the specification:

$$\log(w_{it}) = \mu_i + \lambda^0_j + \lambda^{\mathrm{c}}_j c_i + \lambda^{\mathrm{n}}_j n_i + \varepsilon_{it}, \tag{B.9}$$

where μ_i represent worker-specific intercepts. As in the event studies of Sections B.4, we consider workers of skill types (c_i, n_i) and $(c_{i'}, n_{i'})$ transitioning between firms j and j'. The observed difference between wages for type i workers in the two firms is, on average, equal to:

$$\mathrm{E}_{j'}[\log(w_i)] - \mathrm{E}_j[\log(w_i)] = (\lambda^0_{j'} - \lambda^0_j) + (\lambda^{\mathrm{c}}_{j'} - \lambda^{\mathrm{c}}_j)c_i + (\lambda^{\mathrm{n}}_{j'} - \lambda^{\mathrm{n}}_j)n_i. \tag{B.10}$$

Given this expression, the difference-in-differences for transitioning workers deliver the identification result.

Remark 3. *Define the subset of transitioning worker types such that $c_{i'} \neq c_i$ but $n_{i'} = n_i$. Then, conditioning on this subset, it follows that*

$$\lambda^{\mathrm{c}}_{j'} - \lambda^{\mathrm{c}}_j = \frac{\big(\mathrm{E}_{j'}[\log(w_{i'})] - \mathrm{E}_j[\log(w_{i'})]\big) - \big(\mathrm{E}_{j'}[\log(w_i)] - \mathrm{E}_j[\log(w_i)]\big)}{c_{i'} - c_i} \tag{B.11}$$

Identification of noncognitive returns λ^{n}_j and firm intercepts λ^0_j is similarly obtained from appropriate comparisons of earnings and worker types following firm switches. As explained in Section 3.1, and as common in worker–firm models, this identification is relative to a base firm j' or to the average of all firms (i.e., means of firm parameters are normalized).

C Measurement Error in Skill Measures

A possible concern when using direct skill measures is their reliability. In what follows, we examine the effects of measurement error on the estimates of heterogeneous skills returns. Accounting for measurement error in high-dimensional models presents challenges as noise in each skill might affect estimates of returns from all skills. Since work on measurement error is often limited to single skill impacts, we develop a simple procedure to compare estimates of multiple skill returns under different assumptions on the reliability of skill measures. The approach allows for potential cross-effects from noise in different skills. We correct estimated returns by assuming different combinations of measurement error in skill bundles and obtain informative bounds on the magnitude of skill returns under conventional assumptions about the reliability of skill measures.

Reliability ratios. We define the reliability ratio in each skill ($s \in \{c, n\}$) as:

$$r^s = \frac{\mathrm{sd}(s_i^*)}{\mathrm{sd}(s_i^*) + \mathrm{sd}(u_i^s)},$$

where s_i^* are the true cognitive and noncognitive skills and u_i^s are "classical" i.i.d. measurement errors uncorrelated across individuals and attributes. The latter assumption is partly motivated by the lack of estimates for the correlation of errors across traits. Empirical studies often focus on settings where measurement error in one variable is unrelated to errors in other variables.[10] It is worth noting, however, that our suggested approach (described below) could be explicitly adapted to settings where information is available about the correlation of errors across attributes.

Different estimates of reliability ("reliability ratios") have been suggested for the Swedish skill measures. For example, Grönqvist et al. (2017) employ instrumental variables to correct for noise in skill measures when studying intergenerational mobility. Baseline reliability ratios in their paper are 0.731 for cognitive and 0.498 for noncognitive skills. We adopt $r^c = 75\%$ and $r^n = 50\%$ as lower bounds for reliability. Another prominent study on the Swedish enlistment scores is Lindqvist and Vestman (2011). They find that reliability ratios are 0.868 for cognitive and 0.703 for noncognitive; we use $r^c = 85\%$ and $r^n = 70\%$ as our mid-range estimate.

C.1 Correcting Estimates Using Reliability Ratios

We employ four steps to assess the sensitivity of skill returns estimates to alternative assumptions about the reliability of skill measures.

1. We begin by using the baseline (observed) skill measures. Taking previous studies as a reference, we define reliability ratios denoted as r^c and r^n, respectively, for cognitive and

[10] For example, Lindqvist and Vestman (2011), in their Appendix C, write: "...We further assume that all cross-moments between the true variables and the measurement errors are zero." Grönqvist et al. (2017, Section II) accordingly refer to "classical" measurement error as defined right below.

noncognitive attributes. The reliability ratios indicate the accuracy of the measurements. Higher reliability ratios imply more accurate measurements.

2. To quantify the effects of measurement error, we introduce additional noise to the measured skill values. By doing so, we incrementally lower the reliability of both cognitive and noncognitive measures. This step allows us to assess the sensitivity of model estimates to varying levels of skill measure inaccuracy.

3. Using the newly constructed skill measures, with reduced reliabilities, we estimate the moments of interest. These include variances, covariances, or other moments of the estimated skill returns that we later employ in the analysis.

4. The last objective of the exercise is to project estimates of the moments of interest on the assumed reliabilities. Having established the relationship between moments and reliabilities, we are able to recover the values associated to full accuracy for cognitive and noncognitive measures. The hypothetical full accuracy scenario provides a gauge of the costs of imperfect skill measures as well as a robustness check of our baseline estimates.

C.2 Findings

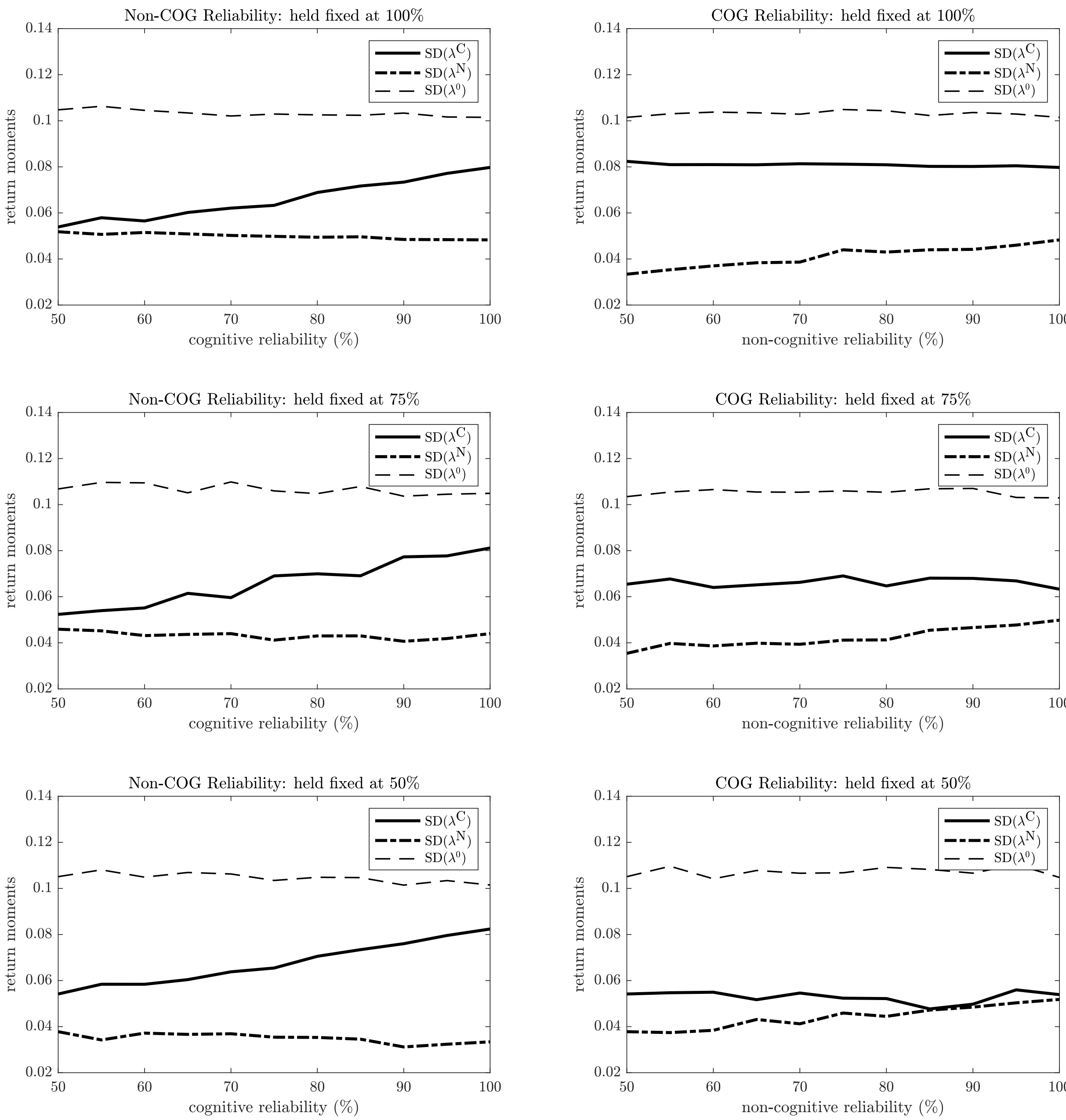
In steps 1-to-3 above, we add measurement error to skill proxies and compute new estimates (with 100 firm groups). To account for possible cross effects, we consider a five-percentage point grid that corresponds to different combinations of reliability ratios $r^c \times r^n \in \{50\%, 55\%, ..., 95\%, 100\%\} \times \{50\%, 55\%, ..., 95\%, 100\%\}$. The 100% reliability values indicate estimates obtained from the original measures, with no additional noise, as reported in the Swedish enlistment scores. We hold the standard deviations of skills fixed by scaling the modified measures. The scaling keeps the data consistent with the convention of the Swedish enlistment agency, which targets a fixed (Stanine) distribution in the population of test takers.

Figure C.1 shows the results of this exercise when we fix noncognitive and cognitive reliabilities in, respectively, the left- and right-hand side panels[11] while varying the reliability of each variable of interest along the grid steps. We find that:

- Moving from right to left on the horizontal dimension of each panel (that is, adding noise) leads to drops in skill returns dispersions that are visible but not extreme. Halving the information in c_i and n_i (moving from 100% to 50% reliability) leads to about 30% reduction in the standard deviations of λ^c and λ^n; this is not unexpected, since we are adding noise to a right-hand-side variable. Moreover, it is interesting that the change in estimated dispersion is approximately linear in reliability. Cross-effects of noise are small since estimates

[11] Results for reliabilities fixed at other levels than 50%, 75%, and 100% are similar.

Figure C.1: Estimated intercept and returns dispersions when varying one skill's reliability while holding the other fixed.



Notes: The figure shows intercept and returns dispersions when classical measurement error is added to cognitive and noncognitive skills. In each panel, one skill dimension's reliability ratio is fixed while the other's is varied from 50 to 100%, where 100% represents the measure as reported in the Swedish enlistment scores without further error added.

for a given skill do not substantively vary with the reliability level of the other (see next bullet).[12]

- In multivariate analysis, biases due to measurement error in one variable could spill over to coefficient estimates in the other.[13] It is not obvious how the intensity of measurement error in one skill measure should affect estimates of skill returns in the other. Unlike previous studies, we explore the *dispersion* of return coefficients using a high-dimensional model and we are not aware of any formal treatment of biases in higher dimensional settings. Given these considerations, we develop a simple procedure to assess the empirical biases introduced by noise in skill proxies. It turns out that, in our analysis, there are little or no cross-effects on the estimated dispersion of returns along the whole reliability grid. As mentioned before, estimated dispersion of returns in each skill varies linearly with the changes in that skill proxy's reliability.

- Since we find no evidence of substantial spillover effects on the dispersion of estimated firm intercepts, we do not report additional details about the dispersion of the λ^0 parameters.

- We also find that adding large amounts of measurement error in a given skill (say, cognitive relative to noncognitive, for example assuming $r^c = 50\%$, $r^n = 100\%$) does not revert the ranking of estimated dispersions in λ^n relative to λ^c. This corroborates our baseline finding that the dispersion of cognitive returns is at least as large as that of the noncognitive. As we show, this result is robust to different combinations of measurement error in skills.

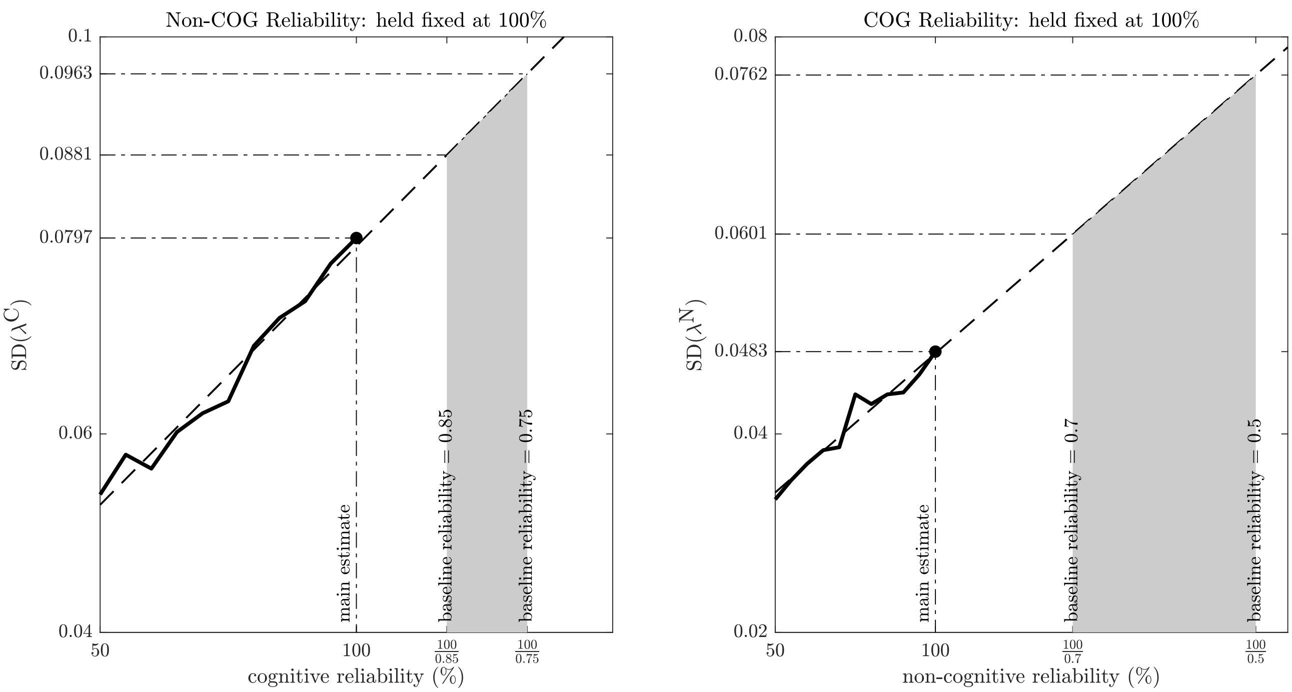
Estimates under the assumption of higher reliability. Estimates of returns' dispersion in each skill change with the reliability of that skill and there is little evidence of spillover effects from error in variables across skill returns. It is, therefore, informative to extrapolate estimated dispersions under the assumption of higher reliability ratios; in particular, this is helpful to assess the extent to which departures from full reliability affect baseline estimates. The exercise is described in step 4 of the procedure outlined above and the results are shown in Figure C.2.

First, we replicate and plot the analysis from the previous steps, showing estimates of returns' dispersions when moving to the left of the baseline estimates (that is, adding more noise). We indicate the baseline estimates by cognitive/noncognitive reliabilities of 100% in the plots. As more noise is added, the $\text{sd}(\lambda^c)$ and $\text{sd}(\lambda^n)$ decline almost linearly from 0.0797 and 0.0483 to about 0.053 and 0.035, respectively.

[12]In the left panels of Figure C.1, moving from 100% to 50% reliability leads to a decline of $\text{sd}(\lambda^c)$ from 0.08 to 0.053, a drop of about 1/3. The extent of the decline and, largely, the level of dispersion do not depend on whether the reliability in λ^n is 50, 75, or 100 percent. In the right panels, moving from 100% to 50% reliability leads to a decline of $\text{sd}(\lambda^n)$ from around 0.05 to 0.035, a decrease of about 30%. This is also independent of the reliability level for the cognitive measure.

[13]E.g., see Lindqvist and Vestman (2011)'s Appendix C: "Since our skill measures are positively correlated (0.388), classical measurement error in one skill measure will imply a bias away from zero for the other skill measure."

Figure C.2: Extrapolation of returns' dispersion, based on reliability values suggested in the literature.



Notes: The figure shows estimates of cognitive (left panel) and noncognitive (right) returns dispersion under different assumptions about measurement error in the Swedish enlistment data. First, the solid black lines to the left of 100% reliability on the x-axes report the dispersion of, respectively, λ^c and λ^n that were also plotted in the top panels of Figure C.1. To the right of the 100% value on the x-axes we show extrapolations from a linear fit through these black lines. The shaded areas of $r^c \in [0.75, 0.85]$ and $r^n \in [0.50, 0.70]$ correspond to the ranges of reliability that are conventionally suggested in the literature (Lindqvist and Vestman, 2011; Grönqvist et al., 2017). The values on the y-axis $\text{sd}(\lambda^c) \in [0.0881, 0.0963], \text{sd}(\lambda^n) \in [0.0601, 0.0762]$ provide the corresponding intervals for the implied (and higher) returns' dispersion.

Then, leveraging the linearity of the relationships, we extrapolate the estimates to the right side of the reliability axis. To determine what level on the x-axis corresponds to full accuracy, we make assumptions about the signal-to-noise content of skill proxies, based on the reliability ratios suggested in existing papers (cited above). For cognitive skills, our range for the skill signal is $[\frac{100}{85}, \frac{100}{75}]$; for noncognitives, it is $[\frac{100}{70}, \frac{100}{50}]$. The "informative" sections of the skill reliability range are indicated by the respective shaded intervals in Figure C.2. On the y-axis, we report the extrapolated estimates in the range of $\text{sd}(\lambda^c) \in [0.0881, 0.0963]$ and $\text{sd}(\lambda^n) \in [0.0601, 0.0762]$. These correspond to even higher magnitudes of skill returns' dispersion than our baseline estimates, which increases the relative importance of the mechanism we study. Since effects are approximately linear, higher reliabilities would result in larger dispersion estimates. That is, our baseline estimates are a lower bound of the skill returns' dispersions.

To summarize, measurement error in skill proxies suggests that baseline estimates of the dispersion of firm-level returns are conservative. The effects of measurement error are, however, not large enough to overturn the main insights of the baseline analysis (e.g., that the dispersion of cognitive returns is at least as large as that of noncognitive returns). Assuming that measurement error is of the magnitude found in the literature, firm-level returns' dispersion would be around 1-to-1.5 percentage points higher for either skill, which amounts to an increase of around 10–20% in the case of cognitives and 20–30% in the case of noncognitives.

D A Labor Market with Two-Sided Heterogeneity and Multi-Dimensional Skills

We examine the interaction of employer and employee heterogeneity within a model featuring workers with different cognitive and noncognitive abilities. We consider a static setting with a continuum of firms, each producing its own distinct product using labor. All firms benefit from more able workers; however, each firm exhibits idiosyncratic returns to skills. Firm-specific skill returns induce sorting of high-skill workers into high-return firms, something that the matching literature has long emphasized. These layers of heterogeneity are embedded in a labor market where employers choose how many workers to hire based on the demand for their output. Equilibrium implies that the labor market clears.

D.1 Production and Market Structure

There is a measure one of workers who differ in their observable cognitive (c) and noncognitive (n) abilities. We let $G(c,n)$ denote the measure describing the distribution of worker types in the economy. A worker's utility from matching with a firm depends on the wage they receive from that firm plus an idiosyncratic preference shock. For worker i, of type (c,n), the utility of working at firm j with wage $w_j(c,n)$ is

$$u_{ij}(c,n) = \beta \log(w_j(c,n)) + \nu_{ij} \tag{D.1}$$

where ν_{ij} captures an idiosyncratic preference for working at firm j. We assume that shocks ν_{ij} are independent draws from a Type I Extreme Value distribution. This specification could be expanded to add firm-level variation in average amenities (Sorkin, 2018).

Workers choose the firms that give them the highest utility. Using standard arguments (McFadden, 1974), the share $q_j(c,n)$ of type (c,n) workers who choose firm j has a logit form

$$\log(q_j(c,n)) = \log(h(c,n)) + \beta \log(w_j(c,n)). \tag{D.2}$$

Equation (D.2) delivers the upward sloping labor supply equation faced by firm j, with elasticity of supply β. The intercept $h(c,n)$ is determined in equilibrium and guarantees market clearing (every worker gets a job), that is

$$h(c,n) = \left[\int w_k(c,n)^\beta \, \mathrm{d}F(k)\right]^{-1} \tag{D.3}$$

where $F(\cdot)$ is the probability measure describing the distribution of firms in the economy.

As in Lise and Robin (2017), the production function is defined at the level of the match and we do not model complementarity between workers within a firm. A worker of type (c,n) employed at firm j produces according to $f_j(c,n)$, where the function f_j describes the output from the firm-worker match. Technology is CRS and a firm's output is the sum of all employees' products.[14] Firm j's total output is

$$y_j = \int f_j(c,n) q_j(c,n) \, \mathrm{d}G(c,n). \tag{D.4}$$

Output market. Each firm j's output is an intermediate input for a final good Y produced by a representative firm through a CES technology. That is, $Y = \left[\sum_{j=1}^{J} \phi_j y_j^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}}$ where $\sigma > 1$ is the elasticity of substitution. Each intermediate's share parameter ϕ_j is the marginal contribution of y_j to output Y and can be interpreted as the output-market power of a firm. Therefore, in the output market, each firm j faces a downward sloping demand curve for its product. Firm j's inverse demand is

$$\log(p_j) = \log(\phi_j) - \frac{1}{\sigma} \log(y_j) \tag{D.5}$$

where p_j is product price, y_j is output, ϕ_j is a firm-specific (inverse) demand intercept, and σ is the output demand elasticity with respect to price.

The intermediate firm's problem. Given output demand and labor supply curves, firm j decides how many workers to hire for each skill type. Firm j's profit maximization problem is:

$$\begin{aligned} \max_{q_j(c,n)} \quad & p_j y_j - \int w_j(c,n) q_j(c,n) \, \mathrm{d}G(c,n) \\ s.t. \quad & y_j = \int f_j(c,n) q_j(c,n) \, \mathrm{d}G(c,n) \\ & \log(p_j) = \log(\phi_j) - \tfrac{1}{\sigma} \log(y_j) \\ & \log(q_j(c,n)) = \log(h(c,n)) + \beta \log(w_j(c,n)) \end{aligned} \tag{D.6}$$

This problem has a closed form solution, with equilibrium wages in firm j

$$w_j(c,n) = \frac{\left(\frac{\beta}{1+\beta}\right)^{\frac{\sigma}{\sigma+\beta}} f_j(c,n) \left(\frac{\sigma-1}{\sigma} \phi_j\right)^{\frac{\sigma}{\sigma+\beta}}}{\left[\int f_j(c,n)^{1+\beta} h(c,n) \, \mathrm{d}G(c,n)\right]^{\frac{1}{\sigma+\beta}}} \tag{D.7}$$

D.2 Base Pay and Skill Premia: Mapping Model to Firm Wages

Firms' production choices can be characterized along the two input dimensions (cognitive and noncognitive). Every worker has a type within the set (c,n), with the first letter denoting cognitive

[14] Additive separability is often assumed in matching models with one-to-many sorting. In the empirical section we show how this technology specification delivers an accurate approximation of returns to different skill types. While convenient, the separability assumption is not crucial for our findings about sorting and returns heterogeneity.

level and the second noncognitive level. The wage premium associated to skill bundle (c,n) in firm j is

$$e^{\Delta_j(c,n)} = \frac{f_j(c,n)}{f_j(L,l)} \tag{D.8}$$

for all (c,n). This corresponds to the wage relative to the low-type worker (L,l), $\left(\frac{w_j(c,n)}{w_j(L,l)}\right)$, since everything else in the wage equation (D.7) cancels. The premium $e^{\Delta_j(c,n)}$ is proportional to the (measurable) productivity of a (c,n) worker in firm j relative to a baseline worker of type (L,l) The parameter $\Delta_j(c,n)$ subsumes two sources of variation: (1) the skill endowment bundle (c,n), and (2) the return to that bundle in firm j. By definition, $\Delta_j(L,l) = 0$ and one can redefine baseline match productivity in firm j as $T_j = f_j(L,l)$, which is the output of workers of type (L,l). Using T_j and $\Delta_j(c,n)$, we write the output of firm j as $y_j = T_j \sum_{(c,n)} e^{\Delta_j(c,n)} q_j(c,n) \mathrm{d}G(c,n)$, where $\mathrm{d}G(c,n)$ with some abuse of notation denotes the total number of (c,n) type workers, and recast the profit maximization as a choice over a discrete set of skill bundles (c,n).

Optimal hiring behavior in the discrete maximization problem implies:

$$w_j(c,n) = \underbrace{\frac{\beta}{1+\beta}}_{\text{Monops.Markdown}} \times \underbrace{\frac{\sigma-1}{\sigma}\phi_j T_j \left(\frac{1}{y_j}\right)^{\frac{1}{\sigma}}}_{\text{Marg.Revenue}} \times \underbrace{e^{\Delta_j(c,n)}}_{\text{Skill Productivity}} \tag{D.9}$$

This expression captures different aspects of market structure. The marginal revenue is an increasing function of a firm's output market share ϕ_j and of its total factor productivity T_j. The latter parameter is normalized to $T_j = f_j(L,l)$, which is the output of a worker with the lowest cognitive and noncognitive ability in firm j. The skill premium $\Delta_j(c,n) = \log(f_j(c,n)/f_j(L,l))$ is the log output in firm j of a (c,n) type worker relative to the lowest skill worker (L,l). the monopsonistic firm sets wages at a fraction $\frac{\beta}{1+\beta}$ of the marginal revenue generated by the worker, with the fraction approaching one in more competitive markets where the labor supply elasticity β is larger. An extra unit of skill rescales marginal revenues proportionally to the firm's skill return $\Delta_j(c,n)$. In log form, the equilibrium wage lends theoretical underpinning the empirical specifications in the paper. That is:

$$\log(w_j(c,n)) = \alpha + \Lambda_j + \Delta_j(c,n). \tag{D.10}$$

The intercept $\alpha \equiv \log\left(\frac{\beta}{1+\beta}\frac{\sigma-1}{\sigma}\right)$ is common across firms and skills, while $\Lambda_j \equiv \log\left(\phi_j T_j y_j^{-\frac{1}{\sigma}}\right)$ is the firm-specific baseline wage, which does not vary with worker skills; $\Delta_j(c,n)$ is a *firm-specific return to skill bundle* (c,n). Under the model's null hypothesis, the firm's demand intercept ϕ_j is subsumed in the fixed effect component Λ_j.

Optimal behavior implies that firms with higher returns to (c,n)-type skills tend to hire a larger share of (c,n)-type workers. This observation suggests that firms with similar returns to a skill type can be grouped together based on their share of workers with that particular type.

E Additional Estimation Results

E.1 Results from Alternative Samples and Estimation Approaches

Columns (1) and (2) in Table E.1 show the standard deviations of individual firm effects when we do not apply the quadratic-form correction (plug-in values) or when the sampling entails leaving single observations (worker–year) out, rather than the whole worker–firm spell as we do in Table 1 of the paper. As expected, when comparing to the baseline results, both these alternative specifications result in more pronounced firm heterogeneity. In this sense, our baseline estimates provide a conservative view of firm return variation. Details about the different sampling approaches (e.g. leaving out one worker-firm observation rather than the whole match) are discussed in the Appendix Section A.3. Columns (3) and (4) in Table E.1 show results when we

Table E.1: Standard deviations of firm parameters in alternative estimations.

	Firm-level (1999–2008):		Grouped (alt. periods):	
	Plug-in	Leave-obs-out	1990–1999	2008–2017
	(1)	(2)	(3)	(4)
sd(λ_j^0)	0.32	0.22	0.10	0.09
sd(λ_j^c)	0.40	0.21	0.10	0.07
$\times$ $90^{th} - 10^{th}$ pct, cog score (c)	0.30	0.15	0.07	0.06
sd(λ_j^n)	0.39	0.17	0.05	0.05
$\times$ $90^{th} - 10^{th}$ pct, noncog score (n)	0.30	0.13	0.04	0.04
$\times$ $90^{th} - 10^{th}$ pct, cumulative (c+n)	0.59	0.28	0.11	0.09
# unique firms	19,085	19,085	20,484	22,079

Notes: The table shows standard deviations of parameters λ_j^0, λ_j^c, and λ_j^n estimating (6) in alternative specifications and periods. Column (1) are plug-in estimates at the firm-level without quadratic-form correction. Column (2) quadratic-form corrects the firm-level variances leaving one observation (i.e., worker in a given year) rather than match (i.e., worker–firm spell) out at a time. Estimation period: 1999–2008. Columns (3) and (4) show the firm-clustered estimates in alternative periods 1990–1999 and 2008–2017. Otherwise notes to Table 1 apply.

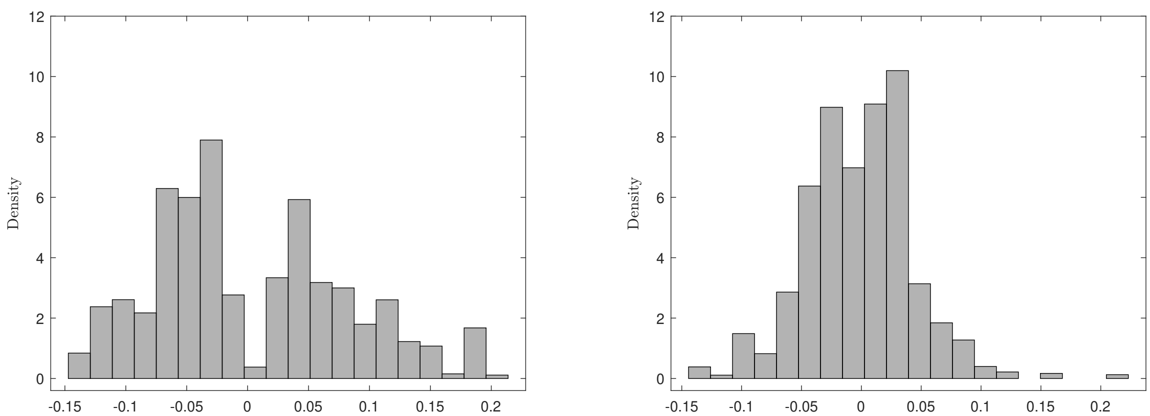
re-estimate the model using the clustering approach for alternative sample periods. Dispersion of firm returns is slightly higher in 1990–1999 than in the baseline estimation period in Table 1. It is slightly lower in 2008–2017, alongside a lower standard deviation of firm intercepts. Overall, the dispersion of firm parameters appears remarkably stable over time.

E.2 The Cross-Section of Skill Returns

To characterize the cross-sectional distribution of firm returns we adopt as a baseline the estimates in Table 1 from the clustered-firms approach. Estimates based on the leave-out bias correction indicate even larger returns heterogeneity.

Figure E.1 shows histograms of cognitive and noncognitive returns in the cross-section of firm clusters. The average λ_j^c and λ_j^n are normalized to zero. Return heterogeneity is significant in both dimensions although larger for cognitive traits, since $\mathrm{sd}(\lambda_j^{\mathrm{c}}) = 0.080$ and $\mathrm{sd}(\lambda_j^{\mathrm{n}}) = 0.048$.

Dispersion is stable across time periods, with $\mathrm{sd}(\lambda_j^{\mathrm{c}}) = 0.095$ and $\mathrm{sd}(\lambda_j^{\mathrm{n}}) = 0.052$ in 1990–1999 and $\mathrm{sd}(\lambda_j^{\mathrm{c}}) = 0.074$ and $\mathrm{sd}(\lambda_j^{\mathrm{n}}) = 0.048$ in 2008–2017 (see Appendix E.1). Edin et al. (2022) show that the *average* return to noncognitive skills increased while that of cognitive skills declined (see also Beaudry et al., 2016; Deming, 2017, for the U.S.).[15] Our analysis suggests that, at the same time, the *heterogeneity* of skill returns across firms did not change differentially for cognitive and noncognitive skills.



(A) Cognitive, $\mathrm{sd}(\lambda^{\mathrm{c}}) = 0.080$

(B) Noncognitive, $\mathrm{sd}(\lambda^{\mathrm{n}}) = 0.048$

Figure E.1: Histograms of Firm Returns (20 Bins)

Notes: Estimates of λ^{c} (left panel) and λ^{n} (right panel), based on 100 firm clusters weighted by employment. $\mathrm{corr}(\lambda^{\mathrm{c}}, \lambda^{\mathrm{n}}) = 0.083$. Grouped estimator for period: 1999–2008.

The employment-weighted correlation of returns among firm clusters, $\mathrm{corr}(\lambda^{\mathrm{c}}, \lambda^{\mathrm{n}})$, is positive at 0.083. Using the firm-level estimates of column (1) in Table 1, the bias-corrected correlation is 0.27. Imperfect correlation lends support to the hypothesis that firm heterogeneity is genuinely multidimensional and that parameters can be independently identified through observable proxies that account for the skill-dependent ranking of workers.

[15] Hermo et al. (2022) show that also returns to cognitive skills' sub-dimensions in Sweden have changed over time.

Earnings gaps and skill premia. The plots in Figure E.1 show that cognitive returns are concentrated between -15 and $+20$ log points. Relative to a worker from the 10th percentile of skills, a worker from the 90th percentile who moves from the bottom to the top of the returns distribution would gain 25 extra log points in earnings. That is, the difference in the cognitive premium between these workers is the skill difference $(0.875 - 0.125 = 0.75)$ multiplied by 35 log points. Complementarity of skills and returns implies that the earning function should be convex over skills because large earning effects accrue from matching high c workers to high λ^{c} firms. Noncognitive returns have a similar range of variation and add significantly to these earning differences. The impact of returns heterogeneity on the distribution of earnings hinges on the intensity of assortative matching. In main text Section 4, we derive testable restrictions to gauge the prevalence of assortative matching in data. Then, in Section 5, we examine how firm heterogeneity, and the responses it elicits, shape the earnings distribution, and contrast our estimates to a counterfactual with random assignment of workers.

E.3 Worker–Firm Matching

Does firm heterogeneity matter for the allocation of workers across employers? And how does it affect the distribution of earnings? To examine these questions we adopt the analytical characterization of matching proposed in Lindenlaub (2017), which describes one-to-one worker–firm matching in a setting with multiple skill attributes. The following results apply to many-to-one settings typically examined in matched employer–employee datasets like ours.

E.3.1 Skill Variation Across Firms

First, we introduce some notation. Firms differ in three dimensions: each firm has a different wage intercept (λ^0) as well as cognitive (λ^{c}) and noncognitive (λ^{n}) returns. Picking up from Section D.1, we call $G(c,n)$ the measure of skills in the working population and let $q_j(c,n)$ be the fraction of the total workforce of type (c,n) hired by firm j. Given the bilinear wage function, labor supply equation (D.2) becomes

$$\log(q_j(c,n)) = \log(h(c,n)) + \beta(\mu + \lambda_j^0 + \lambda_j^{\mathrm{c}} c + \lambda_j^{\mathrm{n}} n), \tag{E.1}$$

where $\beta > 0$ is an elasticity of skill supply to pecuniary returns and $h(c,n)$ captures the relative scarcity of each skill bundle (c,n). The employment relationship (E.1) concisely describes an (empirically-consistent) upward sloping labor supply of type (c,n) to firm j.

Finally, we define Q_j, the total number of workers in firm j, as

$$Q_j = \int q_j(c,n)\mathrm{d}G(c,n).$$

Using this notation, the average cognitive and noncognitive ability of workers in firm j are:

$$\overline{c}_j = \int c\, \frac{h(c,n)e^{\beta(\mu+\lambda_j^0+\lambda_j^{\mathrm{c}}c+\lambda_j^{\mathrm{n}}n)}}{Q_j}\mathrm{d}G(c,n) = \int c\, \mathrm{d}M_j(c,n)$$

$$\overline{n}_j = \int n \, \frac{h(c,n)e^{\beta(\mu+\lambda_j^0+\lambda_j^c c+\lambda_j^n n)}}{Q_j} \mathrm{d}G(c,n) = \int n \, \mathrm{d}M_j(c,n)$$

where M_j is the probability measure of the skills' distribution in firm j. The within-firm measure M_j does not depend on μ or λ_j^0 and only varies with cognitive and noncognitive returns λ_j^c and λ_j^n.

Assortative matching. Assortative matching, whether positive (PAM) or negative (NAM), can be characterized by the properties of the matching function's derivatives. We define matching function $\varphi(\lambda_j^c, \lambda_j^n) = (\overline{c}_j, \overline{n}_j)$, which maps firms' returns into their average worker skills. In matching problems with one dimensional heterogeneity, this boils down to the sign of a single derivative. With multiple attributes, all elements of the Jacobian play a role (see Lindenlaub, 2017).

Definition 1 (PAM as in main text)**.** *The sorting pattern is locally PAM if, for given (λ^c, λ^n), the following holds:*

$$(a)\ \frac{\partial \overline{c}_j}{\partial \lambda_j^c} > 0; \qquad (b)\ \frac{\partial \overline{n}_j}{\partial \lambda_j^n} > 0; \qquad (c)\ \frac{\partial \overline{c}_j}{\partial \lambda_j^c}\frac{\partial \overline{n}_j}{\partial \lambda_j^n} - \frac{\partial \overline{c}_j}{\partial \lambda_j^n}\frac{\partial \overline{n}_j}{\partial \lambda_j^c} > 0.$$

Proposition E.1. *The Jacobian of the matching function evaluated at $(\lambda_j^c, \lambda_j^n)$ is equal to the covariance matrix of the worker-skill distribution within firms with returns $(\lambda_j^c, \lambda_j^n)$. That is,*

$$\frac{\mathrm{d}\varphi(\lambda_j^c, \lambda_j^n)}{\mathrm{d}(\lambda_j^c, \lambda_j^n)} = \beta \, \mathbf{cov}_{M_j}[c,n] = \beta \begin{bmatrix} \mathrm{var}_{M_j}[c] & \mathrm{cov}_{M_j}[c,n] \\ \mathrm{cov}_{M_j}[c,n] & \mathrm{var}_{M_j}[n] \end{bmatrix} \tag{E.2}$$

where the covariance is taken under the within-firm measure M_j.

Proof.

$$\begin{aligned}
\frac{\mathrm{d}\overline{c}_j}{\mathrm{d}\lambda_j^c} &= \frac{\int \beta\, c^2\, h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)}{\int h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)} - \beta \left[\frac{\int c\, h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)}{\int h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)} \right]^2 \\
&= \beta \int c^2 \, \mathrm{d}M_j(c,n) - \beta \left[\int c \, \mathrm{d}M_j(c,n) \right]^2 \\
&= \beta \cdot \mathrm{var}_{M_j}[c]
\end{aligned}$$

$$\begin{aligned}
\frac{\mathrm{d}\overline{c}_j}{\mathrm{d}\lambda_j^n} &= \frac{\int \beta\, cn\, h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)}{\int h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)} \\
&\qquad -\beta \frac{\int ch(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)}{\int h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)} \times \frac{\int nh(c,n)\, e^{\beta(\lambda_j^c n+\lambda_j^n n)}\, \mathrm{d}G(c,n)}{\int h(c,n)\, e^{\beta(\lambda_j^c c+\lambda_j^n n)}\, \mathrm{d}G(c,n)} \\
&= \beta \int cn \, \mathrm{d}M_j(c,n) - \beta \int c \, \mathrm{d}M_j(c,n) \times \int n \, \mathrm{d}M_j(c,n) \\
&= \beta \cdot \mathrm{cov}_{M_j}[c,n]
\end{aligned}$$

Similarly we can show $\frac{\partial \overline{n}_j}{\partial \lambda_j^n} = \beta \cdot \mathrm{var}_{M_j}[n]$ and $\frac{\partial \overline{n}_j}{\partial \lambda_j^c} = \beta \cdot \mathrm{cov}_{M_j}[c,n]$. □

We note two implications of Proposition E.1. First, given a positive elasticity β in (E.1), PAM according to Definition 1 must hold as covariance matrices are positive semi-definite. We test these restrictions in equation (8) and Table 2 of the main text.

Moreover, the matching function with bilinear returns offers a natural analogy to a moment generating function. Further examination of this second property is left for future work.

A different test of sorting. An equivalent definition of positive assortative matching is:

Definition 2 (Alternative formulation of PAM)**.** *The sorting pattern is locally PAM if, for given* $(\bar{c},\bar{n})$*, the following holds:*

$$(a)\ \frac{\partial\lambda_j^{\mathrm{c}}}{\partial\bar{c}_j} > 0; \qquad (b)\ \frac{\partial\lambda_j^{\mathrm{n}}}{\partial\bar{n}_j} > 0; \qquad (c)\ \frac{\partial\lambda_j^{\mathrm{c}}}{\partial\bar{c}_j}\frac{\partial\lambda_j^{\mathrm{n}}}{\partial\bar{n}_j} - \frac{\partial\lambda_j^{\mathrm{n}}}{\partial\bar{c}_j}\frac{\partial\lambda_j^{\mathrm{c}}}{\partial\bar{n}_j} > 0.$$

The Jacobian of the matching function becomes:

$$\frac{\mathrm{d}\varphi(\bar{c}_j,\bar{n}_j)}{\mathrm{d}(\bar{c}_j,\bar{n}_j)} = \begin{bmatrix} \frac{\partial\lambda_j^{\mathrm{c}}}{\partial\bar{c}_j} & \frac{\partial\lambda_j^{\mathrm{n}}}{\partial\bar{c}_j} \\ \frac{\partial\lambda_j^{\mathrm{c}}}{\partial\bar{n}_j} & \frac{\partial\lambda_j^{\mathrm{n}}}{\partial\bar{n}_j} \end{bmatrix} \tag{E.3}$$

Intuitively, it does not matter for the empirical test of PAM whether firms choose workers or vice versa. We therefore consider sorting regressions based on (9) in the main text. These linear forms are similar to the projections of fixed effect onto firm characteristics used in the applied literature (Kline et al., 2020). A strength of this specification is that, under general assumptions, the regression parameters can be correctly estimated from a cross-section of individual non-grouped firms. If returns are measured with error, having λ_j^{c} and λ_j^{n} on the left-hand-side avoids biases in the d-parameters of (9). One can then use these linear projections to test for PAM in the cross-section of individual firms; this is true even if other statistics, such as the R^2, are potentially biased. One caveat is that, while point estimates from these regressions are generally unbiased, standard errors must be corrected for the correlation across the first-stage estimates of the outcome variable (firm parameters).[16]

Table E.2 reports estimates from projections in (9), obtained from non-grouped firm-level data (employees' cognitive and noncognitive skills are averaged into firm-specific $\bar{c}_j$ and $\bar{n}_j$). It is apparent that PAM cannot be rejected since own-partial derivatives and the determinant of the Jacobian are positive throughout. The coefficients on $\bar{c}_j$ for λ_j^{c} are only about half as large as on $\bar{n}_j$ for λ_j^{n}. Flipping this around, $\bar{c}_j$ responds more to a given difference in returns, which again implies stronger sorting on cognitive traits. Results are generally robust to controlling for or weighting by employment size in the different columns of Table E.2.

E.3.2 The Distribution of Skills over Returns and First-Order Stochastic Dominance

A corollary of PAM is that, for each skill attribute, the distribution of higher-skilled workers over firm returns should (first-order) stochastically dominate that of lower-skilled workers. For

[16] We use the correction proposed in equation (7) of Kline et al. (2020) to construct adjusted standard errors.

Table E.2: Projection of Individual Firms' Returns onto their Average Skills.

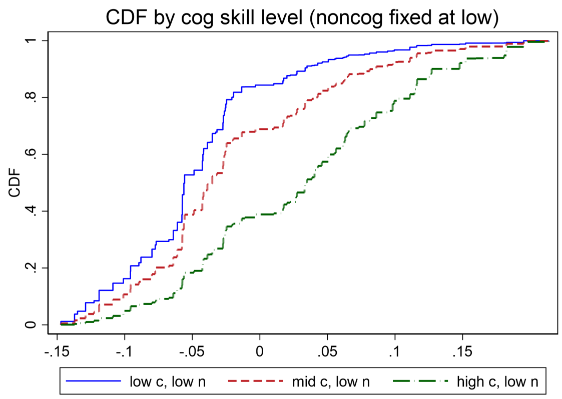
	Dependent Variables:					
	(1)		(2)		(3)	
	λ_j^c	λ_j^n	λ_j^c	λ_j^n	λ_j^c	λ_j^n
$\overline{c}_j$	0.29	-0.41	0.29	–0.41	0.16	–0.44
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.04)
$\overline{n}_j$	0.15	0.61	0.15	0.61	0.40	0.56
	(0.03)	(0.03)	(0.03)	(0.03)	(0.05)	(0.05)
# firms	19,085		19,085		19,085	
Controls	No		# employees		No	
Weights	No		No		# employees	

Notes: The table reports sorting coefficients d_2 and d_3 from estimating equation (9) using firm-level (non-grouped) λ_j^c and λ_j^n. Projections of individual coefficients in estimation period 1999–2008. Standard errors are corrected to account for the first-stage estimates of the outcome variable as in Kline et al. (2020, Section 4).

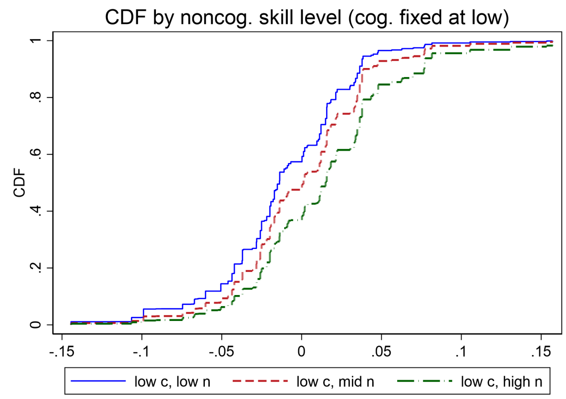
example, holding constant non cognitive traits n, the share of employees with higher cognitive traits c should rise with a firm's λ^c. Equivalently, the frequency of higher n workers should increase with λ^n, holding c constant. Formally, positive FOSD implies:

$$\begin{aligned} &\text{if} \quad c_1 > c_2 \text{ then } \mathrm{CDF^c}(c_1, n, \lambda^c) \leq \mathrm{CDF^c}(c_2, n, \lambda^c) \text{ for all } n, \lambda^c \\ &\text{if} \quad n_1 > n_2 \text{ then } \mathrm{CDF^n}(c, n_1, \lambda^n) \leq \mathrm{CDF^n}(c, n_2, \lambda^n) \text{ for all } c, \lambda^n. \end{aligned}$$

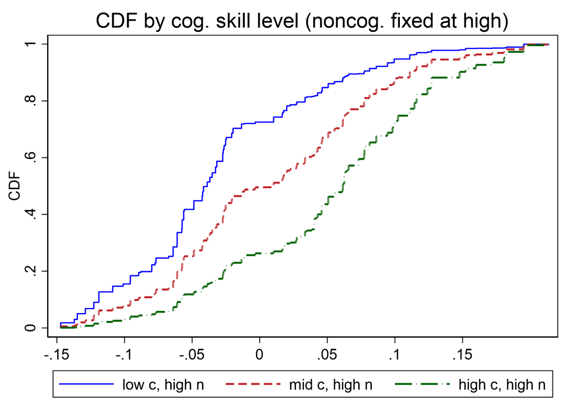
In Figure 2 of the main text we examine whether such patterns occur in our sample. We separately plot the λ^c and λ^n cumulative distribution functions of workers in different skill groups. The top left panel shows that, holding n constant at the medium, the CDF of λ^c shifts to the right when we consider higher endowments of c. A similar finding emerges when looking at the CDF of noncognitive endowments (n) over λ^n (top right panel). Figure E.2 reports the corresponding figures when holding the respective other skill constant at low or high (as opposed to medium) level. All else equal, variation in each skill dimension is consistent with stochastic dominance over the distribution of the corresponding firm returns. This supports the PAM hypothesis and further validates the bilinear matching-return function.



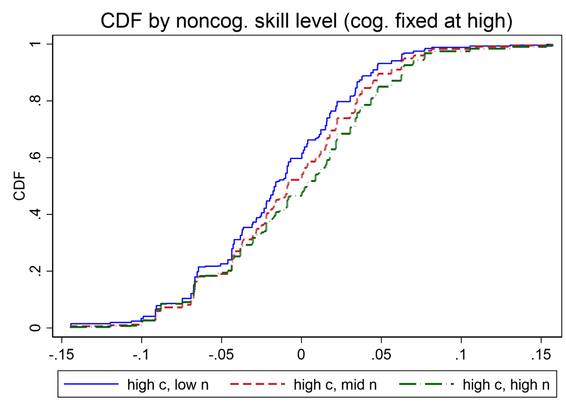
(A) λ_j^c, FOSD (Cognitive)



(B) λ_j^n, FOSD (Noncognitive)



(C) λ_j^c, FOSD (Cognitive)



(D) λ_j^n, FOSD (Noncognitive)

Figure E.2: Distribution of firm returns for different sets of worker skills.

Notes: The figure shows cumulative distribution functions for workers with low ($c,n \leq 0.25$), mid ($0.25 < c,n < 0.75$), or high ($c,n \geq 0.75$) skill ranks over the range of firm returns. Period: 1999–2008. Results from the grouped estimator. FOSD: first-order stochastic dominance.

E.4 Variance Accounting

To facilitate comparisons to existing work, it is useful to characterize the contribution of different layers of firm heterogeneity to the variance of earnings. We perform this exercise for each of the two estimation approaches (bias-correction and clustering). We also report similar decompositions for standard AKM estimators that do not explicitly account for firm-level heterogeneity in skill returns. Finally, to illustrate robustness for each approach, we carry out the analysis for the full sample (where every worker–firm match is observed for possibly multiple periods) and for the collapsed match-level samples (where a worker–firm observation represents the average value over possibly different periods when worker and employer are jointly observed).

Table E.3: Variance decomposition of log earnings (shares × 100). Firms clustered into one hundred classes.

	$\frac{\text{var}(\alpha_{it})}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\psi_{ij})}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\alpha_{it},\psi_{ij})}{\text{var}(log(w_{ij}))}$		
	(1)	(2)	(3)		
Full sample				Obs. (million)	6.48
Full model	60.8	3.8	11.2	total	75.8
AKM	61.0	3.7	11.0	total	75.7
Match-level collapsed sample				Obs. (million)	1.19
Full model	62.0	4.5	14.0	total	80.5
AKM	62.1	4.4	13.9	total	80.4

Notes: Decomposition of the percentage in log earnings variance explained based on estimates from specification (6). We subsume worker-only contributions in $\alpha_{it} \equiv \mu_i + \boldsymbol{X}_{it}\boldsymbol{b}_t$ and firm/worker contributions in $\psi_{ij} \equiv \lambda_j^0 + \lambda_j^c \cdot c_i + \lambda_j^n \cdot n_i$. We group firms into 100 clusters following the approach described in the text. AKM shows results for an alternative model with no heterogeneity in firm returns (here we show fixed effects for the firm groups, not individual firms). Estimation period: 1999–2008.

Notation. To control for variation due to worker-only components, we define $\alpha_{it} \equiv \mu_i + \boldsymbol{X}_{it}\boldsymbol{b}_t$. This is consistent with the normalization outlined in Section 3.1 of the paper, where μ_i encompasses average returns $\kappa_c \cdot c_i + \kappa_n \cdot n_i$ across firms; the α_{it} terms reflect both observed and unobserved worker-level variation. The firm components (including interactions with worker skills) is defined as $\psi_{ij} \equiv \lambda_j^0 + \lambda_j^c \cdot c_i + \lambda_j^n \cdot n_i$. In a standard AKM specification this latter component reduces to firm fixed effects.

Table E.4: Variance decomposition of log earnings (shares $\times$ 100). Variance correction approach: estimates based on sample o individual firms (bias-corrected).

	$\frac{\text{var}(\mu_i)}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\psi_{ij})}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\mu_i,\psi_{ij})}{\text{var}(log(w_{ij}))}$		
	(1)	(2)	(3)		
Leave-one-out sample				Obs. (million)	3.27
Full model	49.4	8.4	5.8	total	63.6
AKM	42.7	8.0	5.3	total	55.9
Match-level collapsed sample				Obs. (million)	1.19
Full model	47.8	7.1	11.0	total	65.9
AKM	42.9	7.6	8.2	total	58.8

Notes: Decomposition of the percentage in log earnings variance explained based on estimates from specification (6). We capture worker-only contributions in μ_i and firm/worker contributions in $\psi_{ij} \equiv \lambda_j^0 + \lambda_j^c \cdot c_i + \lambda_j^n \cdot n_i$. Estimation period: 1999–2008.

Results. Table E.3 shows the variance accounting when we estimate the baseline equation (6) using 100 firm clusters for the 1999–2008 period. Results are similar for the 1990–1999 and 2008–2017 periods, and comparable to other cluster-based implementations for Sweden (Bonhomme et al., 2019) and the U.S. (Lamadon et al., 2022). As in many other studies, worker heterogeneity accounts for much of the total earnings variation while the covariance between α and ψ is the second largest contributor to total variation (see also Bonhomme et al., 2023). One would obtain similar results after restricting the specification to a standard AKM with no skill interactions. This suggests that the significant heterogeneity in skill returns would be mechanically attributed to employer and worker fixed effects. This is of concern when interpreting employer fixed effects as earnings shifts that do not depend on skills.

Table E.4 shows the variance accounting exercise when the coefficients in (6) are estimated using the bias-correction approach. Since this approach adjusts the quadratic forms for worker effects downward, the contribution from worker fixed effects is somewhat lower than the clustering approach although it remains by far the largest. Consistent with the estimates reported in Table 1 of the paper, the direct impact of firm heterogeneity on total variation is larger than in the non-grouped sample. However, due to the downward rescaling of the quadratic forms, the total explained variation is lower than in the clustered estimation. The comparisons to the restricted AKM specifications show that skill returns are mechanically conflated into the employer and worker fixed effects.

Finally, Tables E.5 and E.6 break down the relative contribution of different components of firm effects to total variation. These exercises document that firm level heterogeneity in skill returns accounts for a sizable share (at least $1/4$) of overall firm-specific contributions to inequality and of their covariation with unobserved worker heterogeneity. The latter covariation is quantitatively important if one considers that worker fixed effects include the cross-sectional average of skills returns $\kappa_c \cdot c_i + \kappa_n \cdot n_i$ and account for well over $1/2$ of wage dispersion.

Table E.5: Variance decomposition of log earnings (shares $\times$ 100), including time-varying worker components. Firms clustered into one hundred classes.

	(1)	(2)	(3)		
Variances	$\frac{\text{var}(\alpha_{it})}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\lambda_j^0)}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	Obs. (million)	6.48
	60.8	2.9	0.9		
Covariances	$\frac{2\text{cov}(\alpha_{it},\lambda_j^0)}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\alpha_{it},\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\lambda_j^0,\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	Total explained	75.8
	8.8	2.4	-0.1		

Notes: We group firms into 100 clusters following the clustering approach as described in the text. Estimation period: 1999–2008. We subsume time-invariant and time-varying worker contributions in $\alpha_{it} \equiv \mu_i + \boldsymbol{X}_{it}\boldsymbol{b}_t$.

Table E.6: Variance decomposition of log earnings (shares $\times$ 100), only time-invariant worker components. Firms clustered into one hundred classes.

	(1)	(2)	(3)		
Variances	$\frac{\text{var}(\mu_i)}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\lambda_j^0)}{\text{var}(log(w_{ij}))}$	$\frac{\text{var}(\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	Obs. (million)	6.48
	54.0	2.9	0.9		
Covariances	$\frac{2\text{cov}(\mu_i,\lambda_j^0)}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\mu_i,\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	$\frac{2\text{cov}(\lambda_j^0,\lambda_j^c c_i+\lambda_j^n n_i)}{\text{var}(log(w_{ij}))}$	Total explained	67.8
	7.7	2.4	-0.1		

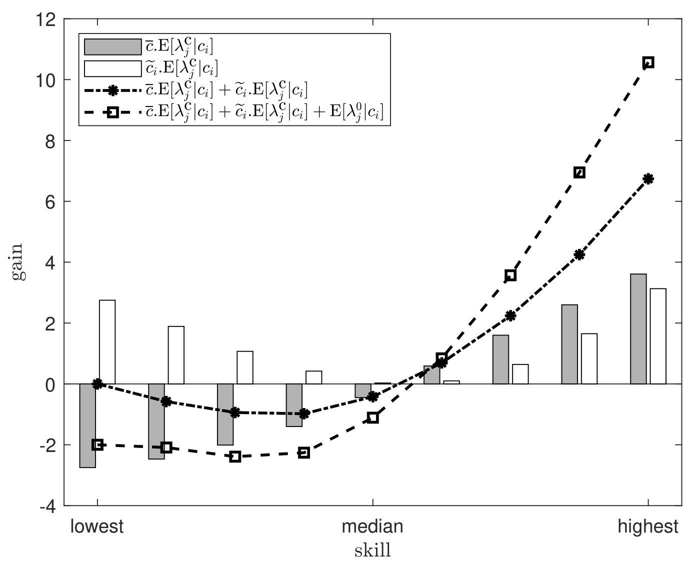
Notes: We group firms into 100 clusters following the clustering approach as described in the text. Estimation period: 1999–2008. We capture time-invariant worker contributions in μ_i.

E.5 The Uneven Gains from Sorting

Robustness to rescaling of skills and returns. The sorting gains discussed in the main text are robust to rescaling of skills and returns since (i) multiplication of c_i by a non-zero factor would lead to a proportional change in the λ_j^c estimates as these would be scaled down by the same factor, leaving the product $\lambda_j^c c_i$ unchanged. (ii) Shifting the level of skills, by adding a constant x to c_i, leaves λ_j^c unchanged and shifts firm intercepts to $\lambda_j^0 - \lambda_j^c x$. Returns from working in firm j become $\lambda_j^c(c_i + x)$ but this is offset by $\lambda_j^0 - \lambda_j^c x$.

The total sorting gain, corresponding to the sum of both intercepts and returns ($\lambda_j^0 + \lambda_j^c c_i$), is hence fully invariant. This cumulative effect, calculated as the sum of columns (2) and (6) in Table 4, induces even larger inequality and skewness across the range of skill levels. Match effects are completely unaffected by rescaling, since they are defined relative to the demeaned $\widetilde{c_i}$.

Figure E.3: Gains from sorting across returns λ_j^c for different cognitive skill levels.



Notes: Gains are multiplied by 100 (i.e., in log points) for readability. All returns are differences relative to a scenario with no heterogeneity in firm returns. Estimates are based on the grouping approach with detailed numbers in Table 4. Sample period: 1999–2008.

The dashed line in Figure E.3 shows the total sorting gain in the cognitive dimension, that is $\mathrm{E}(\lambda_j^0 \mid c_i) + c_i \cdot \mathrm{E}(\lambda_j^c \mid c_i)$. This induces even wider earning differences between skill levels and retains the strong convexity. The average effect, i.e., the aggregate gain from matching, is exactly the same as for the thick dotted line $c_i \cdot \mathrm{E}(\lambda_j^c \mid c_i)$ already seen in the main text.

Table E.7: Gains from sorting across returns λ_j^{n} for different noncognitive skill levels.

	$\mathrm{E}(\lambda_j^{\mathrm{n}} \mid n_i)$ (1)	Full gain (2)	Returns effect (3)	Match effect (4)	$\mathrm{E}(\lambda_j^{0} \mid n_i)$ (5)
skill level (n_i):					
1 (lowest, $n_i = 0$)	-1.52	0.00	-0.78	0.78	-2.76
2	-1.26	-0.16	-0.65	0.49	-2.01
3	-0.87	-0.22	-0.45	0.23	-1.15
4	-0.48	-0.18	-0.25	0.07	-0.04
5 (median, $n_i = 0.5$)	0.06	0.03	0.03	0.00	0.02
6	0.42	0.26	0.22	0.05	0.39
7	0.68	0.51	0.35	0.16	0.91
8	0.84	0.73	0.43	0.30	1.45
9 (highest, $n_i = 1$)	0.94	0.94	0.48	0.45	2.02
Aggregate	0.00	0.13	0.00	0.13	0.00

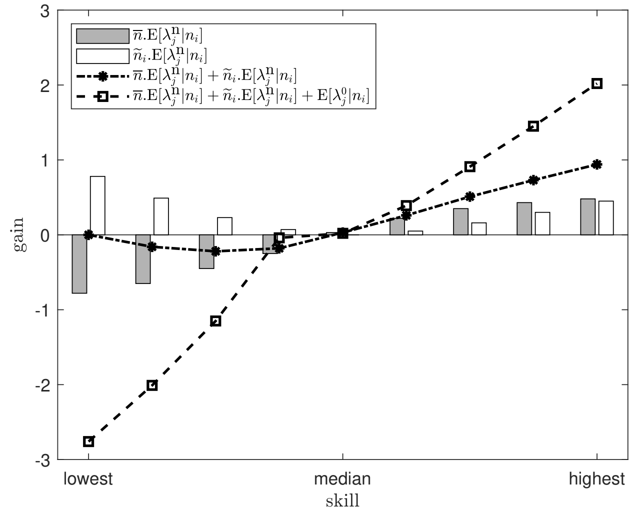
Notes: Gains are multiplied by 100 (i.e., in log points) for readability. All returns are differences relative to a scenario with no heterogeneity in firm returns. Estimates are based on the grouping approach. Sample period: 1999–2008. Column (1): expected marginal return conditional on skill. Column (2): total gain from sorting. Column (3): gain from sorting for the average-skill worker. Column (4): gain from sorting in excess of an average-skill worker with the same employer. Column (5): gain from sorting into intercepts.

Gains from sorting on noncognitive returns. Table E.7 reports the effects from the sorting of noncognitive attributes n_i across noncognitive returns λ_j^{n}. These effects are comparatively smaller than in the cognitive dimension, which reflects the lower dispersion of noncognitive returns across firms (see Section 3) and the weaker sorting in that dimension (see Section 4). Nonetheless, there is clear evidence of sorting also in the noncognitive dimension.

Column (1) in Table E.7 shows that workers with higher noncognitive endowments sample from a distribution of employers with higher returns. Moving from $n_i = 0$ to $n_i = 1$ there is a 2.5 log points difference in $\mathrm{E}(\lambda_j^{\mathrm{n}} \mid n_i)$. This again leads to non-monotonic gains, since high-skill workers benefit the most from the sorting whereas the lowest-skill workers would benefit (or lose) little from any skill returns. The workers who experience steep losses are those with intermediate skills since they would gain from matching with high return firms but are not assigned to such firms. Match effects in column (4) reflect the complementarity of high-skill workers with high-return firms, and of low-skill workers with low-return firms, as well as the induced sorting. These are again positive and raise aggregate earnings by 0.13 log points. In the last column of Table E.7, inequality is further increased by the sorting of noncognitive attributes n_i over λ_j^0 intercepts.

Figure E.4 represents these effects visually. The earnings differences between skill levels are clearly convexified by the sorting (thick dotted line), albeit the convexification is not as pronounced as for cognitive traits. Interestingly, sorting over intercepts reverses this convexification (dashed line), since the least skilled workers face particularly low λ_j^0 (see column (5) of Table E.7). As we emphasized in the main body, and as we see here, the purely redistributive fixed

Figure E.4: Gains from sorting across returns λ_j^{n} for different noncognitive skill levels.



Notes: Gains are multiplied by 100 (i.e., in log points) for readability. All returns are differences relative to a scenario with no heterogeneity in firm returns. Estimates are based on the grouping approach with detailed numbers in Table E.7. Sample period: 1999–2008.

effects (due to sorting into firm intercepts with no complementarity) do not in general induce skewness of the earnings distribution.

F Extensions and Robustness

Firm heterogeneity in skill returns encourages sorting and affects the earnings distribution. One may, however, question to what extent the assignment of workers to jobs occurs along the industry and occupation dimensions. This motivates a robustness exercise where we explicitly test for return heterogeneity within narrowly defined industry and occupation groups.

In addition, and to aid interpretation of our baseline findings, we examine the correlation of skill returns with a subset of firm-level measurements. This is facilitated by external data about firms' balance sheets, capital composition and innovation activities that can be linked to our sample of employers. The latter measures convey information about the nature of production arrangements that may underpin firm differences in skill returns.

Finally, we examine the robustness of estimates under the clustering approach to alternative choices about the number of firm classes and of variables used for grouping firms.

F.1 Industries and Occupations

Using occupation and industry identifiers we can assess whether return heterogeneity is genuinely firm-specific. To this purpose we add industry and occupation interactions with cognitive and noncognitive skills to the specification (6). That is, $\boldsymbol{X}_{it}\boldsymbol{b}_t$ now contains $\lambda_o^c \cdot c + \lambda_o^n \cdot n$ as additional controls where each o indexes one industry or occupation cell.

Table F.1 reports the results, with the first column referring to the baseline specification from the main text for comparison. In column (2) we add industry-specific cognitive and noncognitive skill returns (for 19 different sectors). The contributions of firm intercepts and of returns heterogeneity to earnings dispersion decline very slightly – from 0.10 to 0.09 for sd(λ_j^0) and from 0.06 to 0.05 for sd($\lambda_j^c c_i + \lambda_j^n n_i$). The overall effects remain similar. Column (3) adds detailed five-digit industries, with up to 586 separate returns for each skill dimension; also in this case, the contributions of firm-level parameters to overall dispersion remain stable.

Occupation information is only available in the LISA data from 2001 onward (and only partially before then) so that the estimation sample shrinks. This can be seen, e.g., in the lower number of unique firms in the bottom row of Table F.1.

Introducing occupation-specific returns has more influence on the firm-level parameters. In column (4) of Table F.1 we allow for heterogeneous returns for eight major occupation groups (similar to those used in Acemoglu and Autor, 2011). In this specification the standard deviations of baseline cognitive and noncognitive returns, as well as their contributions to earnings dispersion, decline by about one third compared to the benchmark in column (1). This partly reflects variation in production arrangements within firms; to the extent this variation underpins firm-specific skill returns, it is natural to expect it to be captured by occupation-specific returns. Put differently, the firm-level occupation make-up is one of the primitives accounting for firm heterogeneity in skill returns and, therefore, is a legitimate component of the total firm return. Finer occupations in column (5) and even industry-sector × occupation-group interactions in col-

Table F.1: Dispersion of estimated effects under industry / occupation controls.

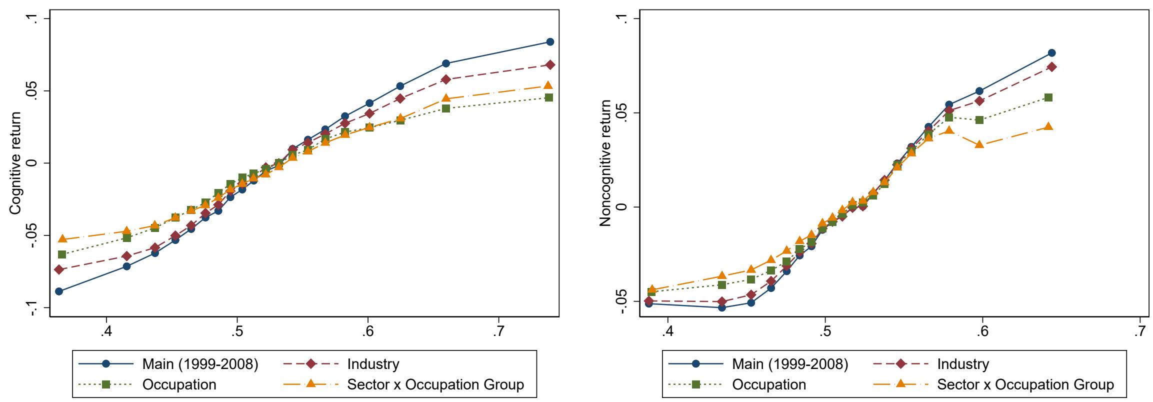
	Main (1)	Sector (2)	Industry (3)	Occup-Group (4)	Occupation (5)	Sec×OccGr (6)
sd(μ_i)	0.43	0.43	0.43	0.41	0.40	0.40
sd(λ_j^0)	0.10	0.09	0.09	0.10	0.09	0.09
sd(λ_j^c)	0.08	0.08	0.07	0.05	0.05	0.05
sd(λ_j^n)	0.05	0.05	0.04	0.04	0.04	0.04
sd($\lambda_j^c c_i$)	0.05	0.05	0.04	0.03	0.03	0.03
sd($\lambda_j^n n_i$)	0.03	0.03	0.03	0.03	0.02	0.02
sd($\lambda_j^c c_i + \lambda_j^n n_i$)	0.06	0.05	0.05	0.04	0.04	0.04
# unique firms	25,783	25,783	25,783	23,999	24,168	23,973

Notes: Parallel to Tables 1 and 3, this table shows standard deviations of worker and firm effects but controlling for industry- or occupation-specific skill returns in equation (6). Column (1) repeats our specification from the main text without such controls. Column (2) adds broad industry sector specific skill returns (19 unique values per skill dimension). Column (3) adds detailed industry specific skill returns (up to 586 unique values per skill dimension). Column (4) adds broad occupation group specific skill returns (8 values, these groups can be seen in Figure F.2). Column (5) adds detailed occupation specific skill returns (113 values). Column (6) adds industry-sector × occupation-group specific skill returns (152 values). Group-level estimates in period: 1999–2008.

umn (6), which proxy for specific jobs in a firm, have little additional effect on the contribution of firm heterogeneity to earnings dispersion.[17]

Sorting patterns. Figure F.1 shows that the patterns of skill sorting across returns are effectively unchanged when we control for industry and occupation-specific interactions. The range of variation of firm-level returns is only slightly smaller, in line with the reduction of dispersion in Table F.1. Sorting across firms remains strong and remarkably robust over the skill range.

[17]While results would not be much different than the industry-sector × occupation-group specification, we refrain from explicitly reporting estimates of detailed industry × detailed occupation-specific returns estimates. The reason is that this has additionally more than 21 thousand nonmissing cell-specific returns (almost as many as there are firms) for each skill dimension and thus reinstates an incidental parameter bias problem that the group-level estimation shown here circumvents.



(A) λ_j^{c} rises in $\overline{c}_j$

(B) λ_j^{n} rises in $\overline{n}_j$

Figure F.1: Average skill by estimated return under different industry / occupation controls.

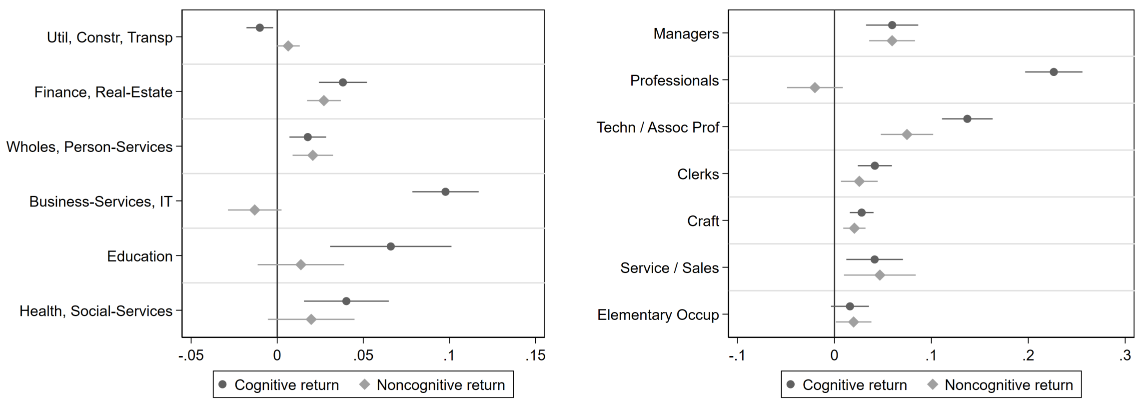
Notes: Parallel to Figure 2, the figure plots binned scatterplots of firm-specific skill returns (vertical axis) with average skills (horizontal axis) for the baseline specification shown in the main text; additionally controlling for detailed industry specific skill returns (up to 586 unique values per skill dimension) in equation (6); controlling for detailed occupation specific skill returns (113 values); and for industry-sector × occupation-group specific skill returns (152 values). Group-level estimates in period: 1999–2008.

We conclude that firm-level differences are an important source of skill return heterogeneity. Accounting for industry and occupation heterogeneity provides further evidence of the large differences that persist at the firm level; these differences do not reflect purely sectoral or occupational variation. Rather, we find that even within the same narrow industries and occupations, skills command significantly different returns across employers.

Aggregating returns to the industry and occupation level. Whereas most of the heterogeneity occurs at the firm level, one may ask which industries or occupations exhibit higher skill returns on average. To answer this question, we first consider linear projections of baseline estimates of λ_j^{c} and λ_j^{n} on a full set of seven industry sector dummies. The projections are similar to those described in equation (9), where $\overline{c}_j$ and $\overline{n}_j$ are replaced by sector dummies, and yield the average cognitive and noncognitive return in the respective industry compared to the omitted "Manufacturing" sector.

Figure F.2A summarizes the results for the group-level estimates in the form of a coefficients plot. Cognitive returns are especially high in the business services and IT sector, noncognitive returns tend to be higher in wholesale and personal service related activities. By contrast, cognitive returns are rather low in the omitted manufacturing sector itself (represented by the zero line) and in utilities, transport, and services. Noncognitive returns in addition are remarkably low in business services and IT.

Figure F.2B shows corresponding results from a linear projection of estimates of λ_j^{c} and λ_j^{n} onto an exhaustive set of employment shares for eight broad occupation groups. The base-



(A) Industry sectors

(B) Occupation shares

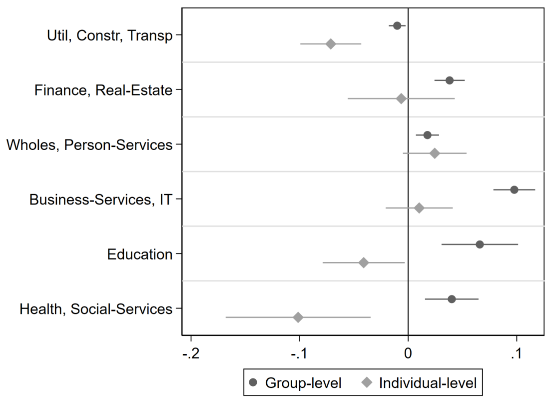
Figure F.2: Skill returns by industry and occupation composition.

Notes. Panel (a): coefficients from the projection of skill returns λ_j^c and λ_j^n onto seven broad industry-sector dummies. Sector dummies add up to one and the omitted sector is "Manufacturing", i.e., coefficients indicate difference in average skill return compared to average in manufacturing ($\lambda_j^c = -0.029$ and $\lambda_j^n = -0.004$ in that sector). Panel (b): coefficients from the projection of λ_j^c and λ_j^n onto a full set of eight broad occupation employment shares in each firm. Occupation group shares sum to one and the omitted group is "Operators / Assemblers". Returns are estimated for 100 firm classes. 95% confidence intervals based on robust standard errors clustered at the level of firm classes.

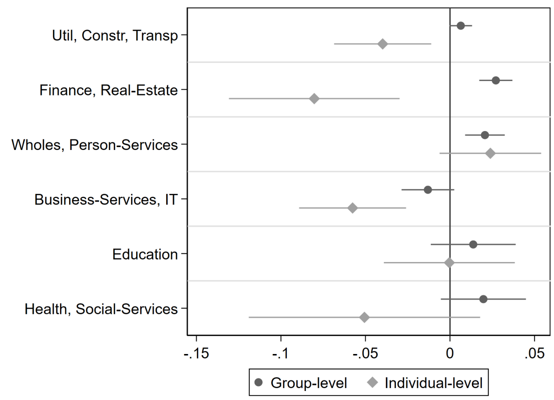
line omitted occupation are "Operators / Assemblers", a large manufacturing-type occupation group. As in Table F.1, and likely because they can vary within firms, occupations are somewhat more related to cognitive and noncognitive returns. That is, firms with large shares of professional, technical, and clerical workers have significantly higher cognitive returns compared to operator/assembler workers. Firms with larger shares of managers, technical workers, and services/sales workers have both high cognitive and noncognitive returns. As for business services and the IT sector, noncognitive skill returns are low among firms with a high share of professional workers.[18]

Finally, results are robust to alternatively considering the firm-level estimates of λ_j^c and λ_j^n from the (smaller) leave-match-out sample. This is shown in Figure F.3, next to the group-level estimates. This approach is less precise and has wider confidence intervals but remains broadly consistent with the group level projections. These exercises suggest that firms in certain industries, and with certain occupations, differentially reward particular skills. Yet, while such variation exists, skill returns (even conditional on, say, a given occupation) vary substantially across firms. In fact, occupation composition can itself be an outcome partly driven by return differences across firms.

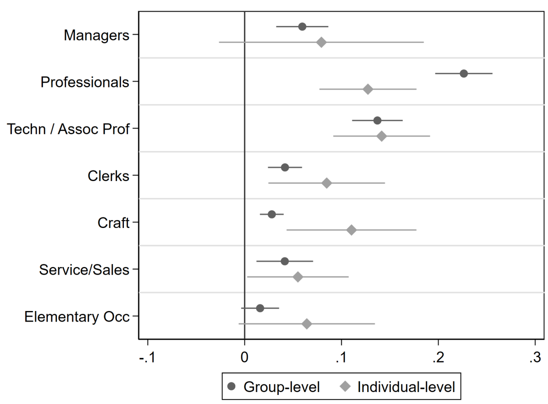
[18]These low noncognitive returns are consistent with the cross-sorting we found in Section 4 if the very high cognitive returns attract very cognitively able professionals to those firms. The professionals also have high noncognitive skills but they accept the low noncognitive returns in the "business / professional services" firms in exchange for the exceptional returns on their cognitives.



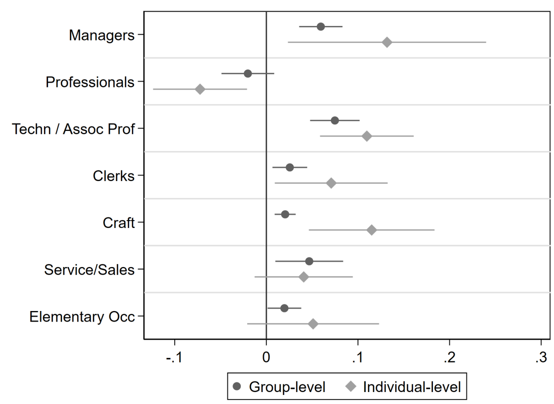
(A) Cognitive λ_j^c, industry sectors



(B) Noncognitive λ_j^n, industry sectors



(C) Cognitive λ_j^c, occupation shares



(D) Noncognitive λ_j^n, occupation shares

Figure F.3: Skill returns by industry and occupation, firm- versus group-level estimates.

Notes. See note to Figure F.2. Here we additionally plot the projections of firm-level λ_j^c and λ_j^n estimates onto broad industry sector dummies and occupational employment shares, and then compare them to the projections of group-level estimates from Figure F.2 separately by skill dimension.

F.2 Capital Composition, Innovation, and Skill Returns

Balance sheets and capital components. We use a commercial data product, the "Serrano" database provided by Bisnode AB, that collects and cleans information about each firm's financials. Up to now, and consistent with prior work, we have referred to workplaces as "firms". However, for the balance sheet analysis we aggregate workplaces up to the organization level (a broader notion of "firms") for which both financial accounts and innovation activity are reported. Since there are multi-workplace corporations, this reduces the number of observations by about one third. Table F.2 reports estimates for projections of cognitive skill returns onto firms' tangible and intangible capital components. To account for zero-value observations for finer capital items in the balance sheets, we use the inverse hyperbolic sine (arcsinh) transform.[19]

Capital composition is strongly associated to cognitive returns. Column (1) of Table F.2 shows that tangible assets vary negatively with skill returns but intangible assets exhibit a strong positive correlation. Column (4) illustrates that the negative relationship holds strong for physical capital (buildings, land, and machinery) and the positive relationship is especially intense for intellectual capital (patents, licenses, and capitalized R&D expenses). The notion that intangible capital and intellectual property are complementary to high skilled labor within a firm is consistent with production arrangements that leverage innovation. Relatively high physical assets and machinery, on the other hand, are more frequent in firms that exhibit lower returns to cognitive skills.

These relationships are robust in several respects: they hold within industry sectors of the economy (columns (2) and (4) of Table F.2) and if we weight with firm employment size (columns (3) and (6)). Tables F.3 and F.4 show that they hold in the leave-out firm-level samples as well as when using dummy indicators (or logs) instead of the arcsinh transformation. Results for noncognitive skills are less pronounced and returns are modestly higher in firms with more physical capital. This lends support to the notion that skills should be modeled separately rather than collapsed into a single index. Perhaps unsurprisingly if one considers production arrangements, firms that employ intangible and intellectual assets have substantially higher cognitive skill returns.[20]

Table F.3 shows the projections of skill returns estimated at the group level onto firm capital components per employee in various robustness specifications. First we employ alternatives to the arcsinh transformation of balance sheet items in the main text. Columns (1) and (2) use dummies, which take the value of one when a firm reports a positive value of the respective capital item as opposed to zero (missing values are still removed). We observe that tangible assets, and in particular physical capital, is significantly negatively associated with cognitive returns whereas intangible assets, and especially intellectual capital, is significantly positively related. As a flip-side of this "extensive margin", we also study the "intensive margin" where we use log

[19] The arcsinh approximates $\log(2x_j) = \log(2) + \log(x_j)$. Estimates are interpreted as semi-elasticities (unit changes) for very small values of the transformed variable x_j, and as elasticities for larger values. See Bellemare and Wichman (2020) and note to Table F.2. Findings are robust to alternative approaches; Table F.3 shows that similar results hold at the intensive margin (log transform of capital items) and at the extensive margin (firms with high cognitive returns are more likely to report nonzero intangible assets).

[20] Even controlling for capital composition in equation (6), or allowing for interactions between capital and skill in parallel to occupation-specific skill returns in Section 6.1, has little impact on the heterogeneity of firm-specific skill returns that we uncover.

Table F.2: Projection of Group Returns onto Firm Capital Composition.

	Dependent variable: $\lambda_j^c \times 100$					
	(1)	(2)	(3)	(4)	(5)	(6)
Tangible assets	-0.83	-0.30	-0.53			
	(0.12)	(0.06)	(0.15)			
Buildings, Land, Machinery				-0.92	-0.35	-0.59
				(0.13)	(0.07)	(0.15)
Other tangible assets				0.13	0.05	0.04
				(0.07)	(0.05)	(0.14)
Intangible assets	1.02	0.63	0.85			
	(0.11)	(0.06)	(0.11)			
Patents, licences, capt. R&D				1.17	0.70	0.72
				(0.12)	(0.07)	(0.14)
Goodwill and other intangibles				0.52	0.36	0.52
				(0.09)	(0.06)	(0.09)
R-squared	0.10	0.33	0.08	0.11	0.33	0.09
Number of firms	14,339	14,339	14,339	14,339	14,339	14,339
Sector fixed effects	No	Yes	No	No	Yes	No
Employment weighted	No	No	Yes	No	No	Yes

Notes: Projections of cognitive skill returns onto capital components per employee, using firms' balance sheets. Tangible fixed assets comprise of buildings and land; machinery and equipment; and other. Intangible fixed assets include capitalized expenditure on research and development; patents, licenses, and concessions; goodwill; and other. All variables are transformed using inverse hyperbolic sine, i.e., $\text{arcsinh}(x_j) = \log\left(x_j + \sqrt{x_j^2 + 1}\right)$. The dependent variable λ_j^c is multiplied by 100. Estimates are based on the sample of clustered firms; period 1999–2008. Robust standard errors clustered at the level of each of the 100 firm groups.

transformations of the balance sheet items. As discussed, the number of non-missing observations now drops and especially so for the detailed distinctions within tangible and intangible assets in column (4). Nonetheless, qualitatively and statistically (as well as in terms of coefficient sizes) the results are comparable to those based on the arcsinh transformation.

Table F.4 shows that the projection results onto capital components are remarkably robust even if we instead use the firm-level estimates of cognitive returns. Finally, we note that the relationships with noncognitive returns are weaker, as shown in columns (5) and (6) of Table F.3. If anything, patents, licenses, and capitalized R&D appear slightly negatively related to noncognitive returns (goodwill and other intangibles positively). Overall, these results are consistent with firms exhibiting heterogeneous production arrangements, whereby capital and employment structure vary substantially and lead to different returns to skill attributes, with the stronger impacts holding in the cognitive skill dimension.

Table F.3: Projection of Group Returns onto Firm Capital Composition.

	Dependent variable: λ_j^c or λ_j^n $\times 100$					
	(1)	(2)	(3)	(4)	(5)	(6)
Tangible assets	-3.69		-1.06		0.14	
	(0.87)		(0.15)		(0.09)	
Buildings, Land, Machinery		-2.86		-0.80		0.07
		(0.81)		(0.21)		(0.10)
Other tangible assets		-2.40		0.25		0.12
		(0.35)		(0.10)		(0.05)
Intangible assets	2.95		0.97		0.03	
	(0.36)		(0.10)		(0.09)	
Patents, licences, capt. R&D		3.49		0.64		-0.24
		(0.44)		(0.12)		(0.10)
Goodwill and other intangibles		1.57		0.53		0.19
		(0.27)		(0.13)		(0.06)
Number of firms	14,339	14,339	5,496	862	14,339	14,339
Dependent variable	λ_j^c	λ_j^c	λ_j^c	λ_j^c	$\boldsymbol{\lambda_j^n}$	$\boldsymbol{\lambda_j^n}$
Independent variables as	**dummy**	**dummy**	**logs**	**logs**	arcsinh	arcsinh

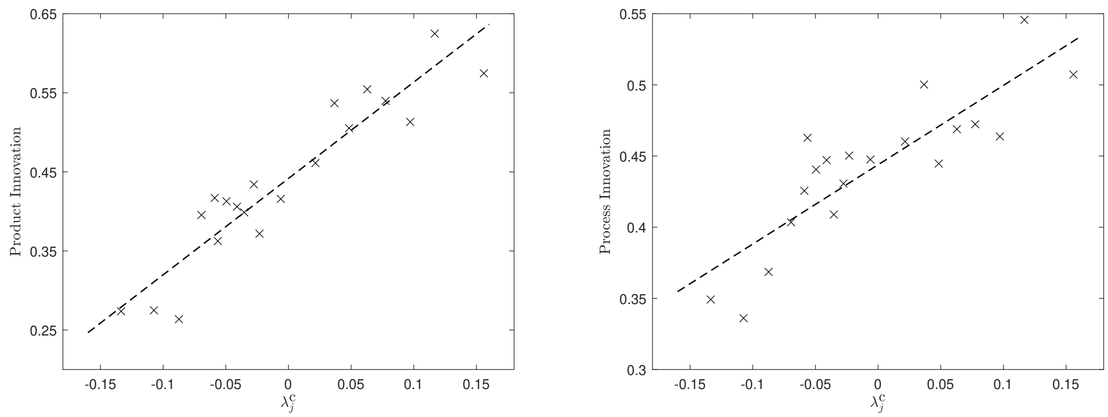
Notes: Results from regressions of skill returns onto capital components per employee from firms' balance sheets. Observations (firms) are unweighted with no further control variables. Columns (1) and (2) use dummies for whether the firm reports a positive value of the respective capital item as opposed to zero. Columns (3) and (4) take logs of the items' values. Columns (5) and (6) use noncognitive instead of the cognitive return as dependent variable with independent variables in arcsinh as in the main text Table F.2. Dependent variables λ_j^c, λ_j^n multiplied by 100. Grouped estimates in period: 1999–2008. Robust standard errors clustered at the level of the 100 firm groups.

Table F.4: Projection of Firm-Level Returns onto Firm Capital Composition.

	Dependent variable: $\lambda_j^c \times 100$ from firm-level estimates					
	(1)	(2)	(3)	(4)	(5)	(6)
Tangible assets	-0.45	-0.56	-0.88			
	(0.36)	(0.41)	(0.19)			
Buildings, Land, Machinery				-0.32	-0.45	-0.59
				(0.44)	(0.46)	(0.25)
Other tangible assets				-0.34	-0.38	-0.55
				(0.40)	(0.41)	(0.22)
Intangible assets	1.03	0.76	0.57			
	(0.32)	(0.33)	(0.16)			
Patents, licences, capt. R&D				1.33	1.03	0.75
				(0.46)	(0.47)	(0.21)
Goodwill and other intangibles				0.46	0.33	0.37
				(0.38)	(0.38)	(0.19)
Number of firms	10,258	10,258	10,258	10,258	10,258	10,258
Sector fixed effects	No	Yes	No	No	Yes	No
Employment weighted	No	No	Yes	No	No	Yes

Notes: Firm-level estimates of λ_j^c in period 1999–2008. Robust standard errors in parentheses. Other than that, see note to Table F.2.

Innovation output. After linking the CIS survey responses to the administrative sample of employers, in Figure F.4 we plot bin scatters of dummies (taking value one in the presence of product/process innovations in the firm) versus cognitive skill returns.[21] Firm innovation activities are positively, and almost linearly, associated with estimates of cognitive returns. This is especially apparent in the case of product innovations where, moving from the lowest to the highest λ_j^c firms, the share of firms which introduce such innovations rises from about 25 to 65 percent. For process innovations the relationship is fainter and only borderline significant when we also condition on product innovation (Table F.5). However, innovation activities still differ by twenty percentage points between firms with the lowest and the highest skill returns.



(A) λ_j^c and product innovation

(B) λ_j^c and process innovation

Figure F.4: Cognitive skill returns and firm innovation.

Notes: The figure plots a binscatter of firms' innovation activities against cognitive skill returns (group-level estimates during 1999–2008). Innovation activities are measured as indicators whether a firm has conducted any product (including service, Panel A) or process (including organizational, Panel B) innovations. This information is from various waves of a representative firm survey (European Community Innovation Survey, CIS). We average the responses (i.e., indicators) for the waves 1998–2000, 2002–2004, 2004–2006, 2006–2008, 2008–2010 relevant to our estimation period. Underlying the plots are 4,138 unique firms. Regression slopes, controlling for a quadratic in firm employment, are $\beta = 1.21$ (clustered S.E. $= 0.13$) and $\beta = 0.55$ (clustered S.E. $= 0.10$) for product and process innovations, respectively.

Similar to the preceding analyses, the first part of Table F.5 shows estimates from projecting cognitive skill returns onto both product and process innovations. As in Figure F.4, we control for a quadratic in employment, since the probability of engaging in innovation rises with the firm's size. The results on product innovations remain strong, whether or not we use group-level (column 1) or firm-level (column 3) return estimates or we control for industry sector fixed effects (i.e., the 19 unique ones from Section F.1).

[21] We plot the raw relationship after controlling for (a quadratic in) employment, since the probability of engaging in any innovation rises with a firm's size. The controls do not substantively affect results. The corresponding relationships for noncognitives are weaker and reported in Appendix F.2.

Table F.5: Projection of Skill Returns onto Firm Innovation Activities.

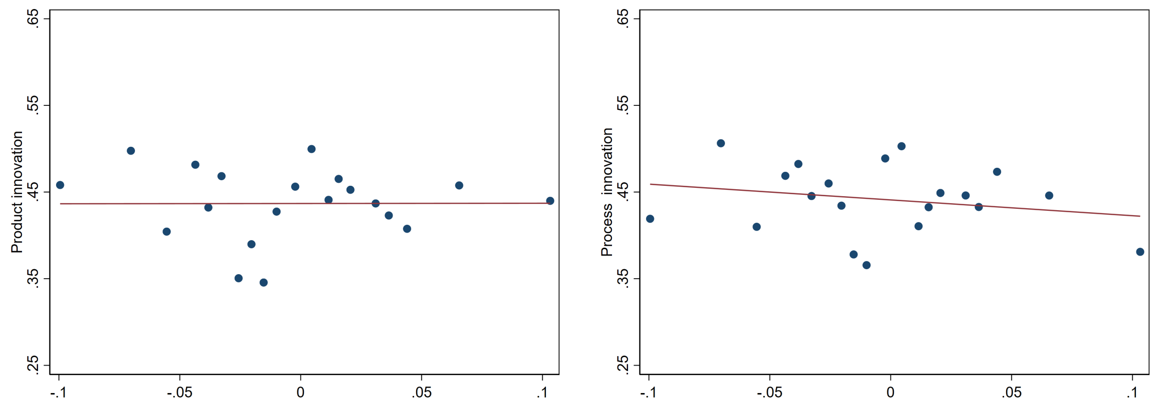
	Dependent variable: $\lambda_j^c \times 100$					
	(1)	(2)	(3)	(4)	(5)	(6)
Innovation output:						
Product innovation	3.73	2.89	7.77			
	(0.53)	(0.37)	(2.61)			
Process innovation	0.29	0.48	4.09			
	(0.30)	(0.25)	(2.73)			
Innovation spending:						
Intramural R&D				0.39	0.30	0.46
				(0.06)	(0.04)	(0.27)
Extramural R&D				-0.01	0.06	0.03
				(0.04)	(0.03)	(0.32)
Acquisition of machinery				-0.14	-0.02	0.05
				(0.04)	(0.03)	(0.27)
Other external knowledge				0.14	0.06	0.37
				(0.04)	(0.03)	(0.31)
Number of firms	4,138	4,138	3,344	3,857	3,857	3,123
Sector fixed effects	No	Yes	No	No	Yes	No
Estimates (level)	Group	Group	Firm	Group	Group	Firm

Notes: The first three columns report estimates from regressions of cognitive skill returns onto indicators for product and process innovations (as defined in the text and note to Figure F.4), controlling for a quadratic in firm employment size. Column (1) uses group-level returns estimates, column (2) adds industry sector fixed effects, and column (3) uses firm-level returns estimates. The last three columns regress returns onto firms' innovation expenditure items, which are $\text{arcsinh}(x_j) = \log\left(x_j + \sqrt{x_j^2 + 1}\right)$ transformed. Otherwise specifications (4)–(6) are parallel to (1)–(3). Returns estimated in period 1999–2008. Robust standard errors in parentheses and clustered at the level of the 100 firm classes for the grouped estimates.

The relationship between skill returns and process innovations gets weaker when we condition on product innovations, and it is only borderline significant.

Specific innovation activities. Next, we examine firms' CIS-reported expenditures on specific types of innovation activities. This is, again, done by using the arcsinh transformation. Column (4) of Table F.5 shows that, consistent with the preceding findings, high cognitive returns firms spend significantly more on intramural (or in-house) research and development. They also spend somewhat more on purchasing external knowledge, and somewhat less on specific machinery. These findings are robust to adding industry sector fixed effects or using estimates of firm-level returns in columns (5) and (6).

Lastly, Figure F.5 shows the baseline binned-scatter plot of skill returns vis-a-vis product and process innovations for the noncognitives. Broadly in line with our prior findings, there is no



(A) λ_j^{n} and product innovation

(B) λ_j^{n} and process innovation

Figure F.5: Noncognitive skill returns and firm innovation.

Notes: The figure plots a binscatter of firms' innovation activities against noncognitive skill returns (group-level estimates during 1999–2008). Innovation activities are measured as indicators whether a firm has conducted any product (including service, Panel A) or process (including organizational, Panel B) innovations. This information is from various waves of a representative firm survey (European Community Innovation Survey, CIS). We average the responses (i.e., indicators) for the waves 1998–2000, 2002–2004, 2004–2006, 2006–2008, 2008–2010 relevant to our sample period. Underlying the plots are 4,138 unique firms. Regression slopes, controlling for a quadratic in firm employment, are $\beta = 0.00$ (clustered S.E. $= 0.32$) and $\beta = -0.18$ (clustered S.E. $= 0.19$) for product and process innovation, respectively.

detectable relationship and, in contrast to λ_j^{c}s, the λ_j^{n}s do not actually predict higher innovation activity.

F.3 Clustering Strategies and Number of Firm Clusters

When using estimators based on firm clusters, one question is whether results are sensitive to the grouping strategy. In what follows, we document how the dispersion (standard deviations) of firm-level parameters, and their contributions to earnings dispersion, vary under alternative restrictions on the number of clusters as well as on the observables used for the clustering.

Table F.6: Alternative clustering specifications.

	Main (1)	10 clusters only (2)	Adding variables (3)	Earning dist. only (4)
sd(μ_i)	0.43	0.44	0.43	0.43
sd(λ_j^0)	0.10	0.07	0.10	0.11
sd(λ_j^c)	0.08	0.06	0.07	0.06
sd(λ_j^n)	0.05	0.03	0.04	0.03
sd($\lambda_j^c c_i$)	0.05	0.04	0.04	0.04
sd($\lambda_j^n n_i$)	0.03	0.02	0.02	0.02
sd($\lambda_j^c c_i + \lambda_j^n n_i$)	0.06	0.04	0.05	0.05
# unique firms	25,783	25,783	25,711	25,783

Notes: Adding to the evidence in Tables 1 and 3, this table shows standard deviations of worker and firm effects under alternative clustering specifications. Column (1) repeats the baseline specification from the main text, for comparison. Column (2) shows estimates when using the means of earnings, cognitive and noncognitive skills within each firm but just ten clusters. Column (3) shows results for 100 clusters after adding standard deviations of earnings, cognitive and noncognitive skills and firm employment size as additional clustering variables. Column (4) shows results when we only use quantiles of the earnings distribution (10th, 30th, 50th, 70th, and 90th) within each firm for clustering and we impose 100 groups. All group-level estimates are based on the sample period: 1999–2008.

Table F.6 reports key results. For comparison column (1) replicates the baseline specification in the main text. In column (2) we use only 10 (rather than the baseline 100) firm clusters; this number is the same as in the main analyses of Bonhomme et al. (2019); Lamadon et al. (2022). The contributions of firm intercepts and skill return heterogeneity to earnings dispersion marginally declines, while the *relative* contribution of returns rises. When we use a richer set of clustering variables – including firm employment size as well as the standard deviations of earnings and cognitive and noncognitive skills, in addition to the means of these variables – re-

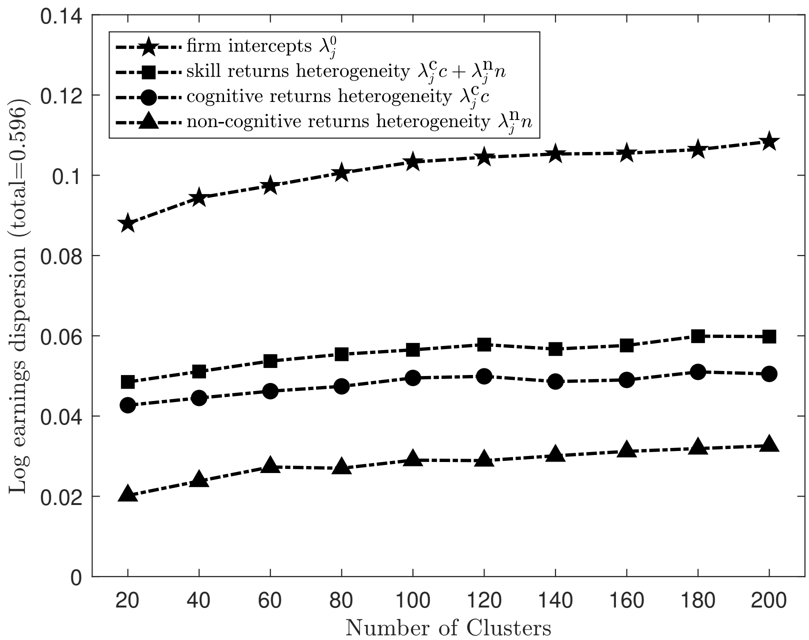


Figure F.6: Dispersion due to firm heterogeneity (log earnings), by number of k-means groups.

Notes: The figure shows the earnings variation due to firm intercepts $\mathrm{sd}(\lambda^0)$, cognitive skill returns $\mathrm{sd}(\lambda_j^c c_i)$, noncognitive skill returns $\mathrm{sd}(\lambda_j^n n_i)$, and overall skill returns $\mathrm{sd}(\lambda_j^c c_i + \lambda_j^n n_i)$ when we re-estimate the model with different numbers of k-means clusters. Estimation period: 1999–2008.

turns heterogeneity does not change significantly either in relative or absolute terms, as shown in column (3). Finally, column (4) shows that restricting the clustering strategy to earnings alone, through the use of quantiles of the earnings distribution within each firm (see Bonhomme et al., 2019; Lamadon et al., 2022), does not materially change the estimated impact of returns heterogeneity when compared to the other robustness checks. In all alternative specifications we also confirm the presence of positive assortative matching patterns.

Finally, Figure F.6 shows estimates of the impact of firm heterogeneity under alternative numbers of clusters (which we let free to vary between 20 and 200). The relative contributions of intercepts and skill returns change only marginally, lending further support to the results obtained under the baseline cluster design.

References

ABOWD, J. M., R. H. CREECY, F. KRAMARZ, ET AL. (2002): "Computing person and firm effects using linked longitudinal employer-employee data," Tech. rep., Center for Economic Studies, US Census Bureau.

ACEMOGLU, D. AND D. AUTOR (2011): "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter and D. Card, Elsevier, vol. 4, Part B, 1043 – 1171.

ACHLIOPTAS, D. (2003): "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of computer and System Sciences*, 66, 671–687.

AGHION, P., U. AKCIGIT, A. HYYTINEN, AND O. TOIVANEN (2023): "Parental education and invention: the Finnish enigma," *International Economic Review*, 64, 453–490.

ANDREWS, M. J., L. GILL, T. SCHANK, AND R. UPWARD (2008): "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 673–697.

BAGGER, J. AND R. LENTZ (2019): "An empirical model of wage dispersion with sorting," *The Review of Economic Studies*, 86, 153–190.

BARTOLUCCI, C., F. DEVICIENTI, AND I. MONZÓN (2018): "Identifying sorting in practice," *American Economic Journal: Applied Economics*, 10, 408–38.

BEAUDRY, P., D. A. GREEN, AND B. M. SAND (2016): "The Great Reversal in the Demand for Skill and Cognitive Tasks," *Journal of Labor Economics*, 34, 199–247.

BELLEMARE, M. F. AND C. J. WICHMAN (2020): "Elasticities and the inverse hyperbolic sine transformation," *Oxford Bulletin of Economics and Statistics*, 82, 50–61.

BÖHM, M. J., D. METZGER, AND P. STRÖMBERG (2023): "Since You're So Rich, You Must Be Really Smart: Talent, Rent Sharing, and the Finance Wage Premium," *Review of Economic Studies*, 90, 2215–2260.

BONHOMME, S., K. HOLZHEU, T. LAMADON, E. MANRESA, M. MOGSTAD, AND B. SETZLER (2023): "How Much Should we Trust Estimates of Firm Effects and Worker Sorting?" *Journal of Labor Economics*, 41, 291–322.

BONHOMME, S., T. LAMADON, AND E. MANRESA (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

——— (2022): "Discretizing unobserved heterogeneity," *Econometrica*, 90, 625–643.

BONHOMME, S. AND E. MANRESA (2015): "Grouped patterns of heterogeneity in panel data," *Econometrica*, 83, 1147–1184.

CARD, D., A. R. CARDOSO, AND P. KLINE (2016): "Bargaining, sorting, and the gender wage gap: Quantifying the impact of firms on the relative pay of women," *The Quarterly Journal of Economics*, 131, 633–686.

CARD, D., J. HEINING, AND P. KLINE (2013): "Workplace Heterogeneity and the Rise of West German Wage Inequality*," *The Quarterly Journal of Economics*, 128, 967–1015.

DEMING, D. J. (2017): "The growing importance of social skills in the labor market," *The Quarterly Journal of Economics*, 132, 1593–1640.

EDIN, P.-A. AND P. FREDRIKSSON (2000): "Linda-longitudinal individual data for Sweden," Tech. rep., Working Paper, Department of Economics, Uppsala University.

EDIN, P.-A., P. FREDRIKSSON, M. NYBOM, AND B. ÖCKERT (2022): "The Rising Return to Noncognitive Skill," *American Economic Journal: Applied Economics*.

EECKHOUT, J. AND P. KIRCHER (2011): "Identifying Sorting - In Theory," *The Review of Economic Studies*, 78, 872–906.

FREDRIKSSON, P., L. HENSVIK, AND O. N. SKANS (2018): "Mismatch of talent: Evidence on match quality, entry wages, and job mobility," *American Economic Review*, 108, 3303–38.

GRÖNQVIST, E., B. ÖCKERT, AND J. VLACHOS (2017): "The intergenerational transmission of cognitive and noncognitive abilities," *Journal of Human Resources*, 52, 887–918.

HAGEDORN, M., T. H. LAW, AND I. MANOVSKII (2017): "Identifying equilibrium models of labor market sorting," *Econometrica*, 85, 29–65.

HANSEN, K. Y., J. J. HECKMAN, AND K. J. MULLEN (2004): "The effect of schooling and ability on achievement test scores," *Journal of Econometrics*, 121, 39–98.

HECKMAN, J. J., J. STIXRUD, AND S. URZUA (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 24, 411–481.

HERMO, S., M. PÄÄLLYSAHO, D. SEIM, AND J. M. SHAPIRO (2022): "Labor Market Returns and the Evolution of Cognitive Skills: Theory and Evidence," *The Quarterly Journal of Economics*, 137, 2309–2361.

JOCHMANS, K. AND M. WEIDNER (2019): "Fixed-Effect Regressions on Network Data," *Econometrica*, 87, 1543–1560.

KLINE, P., R. SAGGIO, AND M. SØLVSTEN (2020): "Leave-out estimation of variance components," *Econometrica*, 88, 1859–1898.

LAMADON, T., M. MOGSTAD, AND B. SETZLER (2022): "Imperfect Competition, Compensating Differentials, and Rent Sharing in the US Labor Market," *American Economic Review*, 112, 169–212.

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

LINDQVIST, E. AND R. VESTMAN (2011): "The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment," *American Economic Journal: Applied Economics*, 3, 101–128.

LISE, J. AND J.-M. ROBIN (2017): "The macrodynamics of sorting between workers and firms," *American Economic Review*, 107, 1104–35.

MCFADDEN, D. (1974): "Conditional logit analysis of qualitative choice behavior," *Frontiers in Econometrics*, 105–142.

SORKIN, I. (2018): "Ranking firms using revealed preference," *Quarterly Journal of Economics*, 133, 1331–1393.