Skill Distance Between Occupations and Post-Training Professional Transitions of Jobseekers

Kevin Michael Frick*, Yagan Hazard, Damien Mayaux*** and Thomas Zuber******

Abstract – Does vocational training help correct structural imbalances in the labour market? We propose a new measure of the skills distance between occupations, obtained by fine-tuning a large language model on a sample of job offers. Using this method, we demonstrate that the "return to employment" differential between jobseekers with and without training is driven by a reallocation of workers towards occupations that are very different from their previous posts in terms of the skills required. From a purely reallocative perspective, however, the return to employment differential associated with vocational training does not appear to be driven by more jobseekers moving to occupations where employers are struggling to recruit.

JEL: J62, J68, J24
Keywords: structural imbalance, training, skills

**Toulouse School of Economics; **Collegio Carlo Alberto and ESOMAS; ***Paris School of Economics; ****Banque de France.*
Correspondence: *thomas.zuber@banque-france.fr*

The opinions expressed in this study do not represent the official positions of the Banque de France, nor of Eurosystem. This research has benefited from the financial support of DARES, as part of the "Training jobseekers – Wave 3" call for research projects. The project also received a government grant via France's National Research Agency (ANR), as part of the Future Investments programme ANR-17-EURE-001; we were also able to make use of the computing capacities of IDRIS thanks to resource allocation scheme 2022-AD011013484R1 managed by GENCI. We would like to thank Cyril Verluise for his insightful advice in the early days of this project, and Pauline Givord and Anne Bucher for their numerous comments and their help with data access, as well as all of those who contributed to the various project steering meetings.

Received in March 2024, accepted in September 2025. Translated from "Distance entre métiers et transitions professionnelles des demandeurs d'emploi après une formation".

Citation: Frick, K. M., Hazard, Y., Mayaux, D. & Zuber, T. (2025). Skill Distance Between Occupations and Post-Training Professional Transitions of Jobseekers. *Economie et Statistique / Economics and Statistics*, 547, 49–67. doi: 10.24187/ecostat.2025.547.2138

Structural imbalances between supply and demand on the labour market have the effect of increasing both the level and the stubborn persistence of unemployment. Among these imbalances, one frequently discussed hypothesis is that of a *mismatch* between workers' existing skills and the skills sought by businesses. This mismatch leads to the coexistence of residual pockets of unemployment and so-called "bottleneck" markets where businesses looking to recruit find themselves frustrated by a scarcity of qualified labour. In theory, these imbalances could be resolved by redirecting the supply of available labour towards those areas of the market experiencing recruitment difficulties, but in practice numerous factors impede such professional mobility.

Vocational training programmes targeting jobseekers aim to address these imbalances. They do so not only by facilitating access to training for jobseekers, but also by (as much as possible) steering existing training resources towards bottleneck occupations, that is, those with the most acute recruitment difficulties. The combined effect of these two aspects should in theory maximise the impact of vocational training in terms of getting people back to work. In France, the skills investment strategy (*Plan d'investissement dans les compétences*, PIC) launched in 2018 was designed with this goal in mind.

The direct effects of training on return to employment are better understood than the indirect effects of training on the balance of the labour market through occupational transitions. Numerous studies have documented the impact of training on return to employment (see for example the meta-analysis proposed by Card *et al.*, 2018). However, with a few exceptions, little academic attention has been devoted to the indirect effects that vocational training can have on the labour market by helping to reduce structural imbalances between the supply and demand of labour in different occupations (Şahin *et al.*, 2014; Barnichon & Figura, 2015; Marinescu & Rathelot, 2018). It remains particularly difficult to quantify the extent to which a given factor, such as geographic mobility, or the skills gap between occupations, can be considered responsible for the imbalances between occupations. Nevertheless, the stakes are high: in France in 2021, these structural imbalances were estimated to be responsible for around 15% of recorded unemployment (Fontaine & Rathelot, 2022).

This article investigates whether, and if so to what extent, jobseekers who have undertaken vocational training go on to make professional transitions *(i)* involving a more significant shift in skills, compared with those who have not retrained, and *(ii)* towards sectors of the labour market where the labour demand is stronger than in sectors in which they would have been able to find employment without further training. To do so, we construct a measure of the distance between occupations in terms of their skill requirements, using the text of job offers published by Pôle Emploi (the French public employment service). Our proposed methodology involves training a neural network on so-called "pretext" tasks so that the model captures content specifically relevant to skills in the corpus of job offers, mapping each offer and each occupation to a low-dimensional skills space and defining the skill distance between any two occupations as the angular distance between the corresponding points within this space.[1] The skills which constitute the dimensions of the space cannot be interpreted individually, but by construction it is possible to compare the skills vector for any two given occupations. By measuring the skill distance between occupations, we can distinguish between professional transitions between occupations demanding relatively similar skills from professional transitions to occupations whose required skills are different from those required by their previous posts. Dawson *et al.* (2021) utilise a similar methodology.[2]

The matching of administrative data provided by the DARES ForCE programme (Training, Unemployment and Employment, French *Formation, Chômage et Emploi*) allows us to track the trajectories of jobseekers who undertook vocational training between 2018 and 2020. Our study focuses on a sub-sample of jobseekers who held a stable job within the year preceding their training programme, thus satisfying the definitions of original occupation and professional transition. Comparing the professional trajectories of trained and untrained jobseekers, leveraging the large number of control variables found in administrative data, we analyse the relationship between training programmes and return to employment, the skill distance between occupations covered during the professional transitions and the resulting differentials in labour market tightness between the occupations involved in post-training professional transitions. We conducted this comparison using double/debiased machine

1. A pretext task is a task used not because it is pertinent to the final training objective for the algorithm, but because it forces the algorithm to acquire certain desirable characteristics.

2. See also Bana et al. *(2019) and Gentzkow* et al. *(2019).*

learning (Chernozhukov *et al.*, 2018) to correct for observable differences between the test group and the control group.[3] The result of this comparison can only be interpreted as a causal effect of vocational training under a conditional independence assumption (CIA) under which selection into vocational training is only driven by observable variables. This is unlikely to be true in practice. In spite of this pitfall, which is amply discussed in the literature on the evaluation of training policies, our results allow us to offer some initial insight into the under-studied connection between vocational training and efforts to address structural imbalances in the labour market.

We study the impact of training on the professional transitions of jobseekers, and certain characteristics of the jobs they find. We can thus show that the relationship between training and the return to employment almost always involves people taking up posts which are very different from their previous occupations. About the reallocation of employment, jobseekers who have retrained are more likely to undertake a professional transition, but it appears that they do not systematically move towards occupations where recruitment difficulties are more acute than they were in their original occupations. This suggests that rethinking the range of vocational training available to jobseekers, to focus more explicitly on those occupations where employers are struggling to recruit enough workers, would serve to improve the results of such training programmes in terms of job prospects. Our methodology requires us to focus exclusively on jobseekers who have held a stable job within the preceding twelve months. This sub-population is younger, more qualified and more likely to find employment than the average jobseeker registered with Pôle Emploi. In addition to the doubts over the validity of our conditional independence assumption, this necessary limitation to our sample also limits the interpretative scope of our results.

This article also contributes to the abundant literature showing that transitioning between occupations comes at a high cost in terms of human capital (Becker, 1964). Shaw (1984) was the first to illustrate the specific importance of career trajectories involving transition between different occupations, rather than different sectors, in terms of determining workers' earning potential. In the context of the German education system, Eckardt (2022) has shown that people working in fields for which they have not been specifically trained earn less money than colleagues with the appropriate qualifications. The cost in terms of lost earnings is positively correlated with the distance between occupations in terms of their skills requirements. The costs associated with forced professional transitions have also been studied, with efforts made to assess the effects of international competition on local labour markets. Traiberman (2019) for Denmark and Basco *et al.* (2025) for France have shown that when people change occupations as a result of strong international competition, the further the new occupation is from their previous occupation, the more money they lose. Moreover, Hyman (2018) has demonstrated that training policies can help workers to make professional transitions towards occupations less vulnerable to international competition. Finally, building upon Shaw's (1984) original intuition, there is now an abundant literature demonstrating that taking the multidimensional nature of skills into account (beyond a simple linear index) is crucial to understand transitions in the labour market, the matching process between candidates and employers, and wage determination (Gathmann & Schönberg, 2010; Lindenlaub & Postel-Vinay, 2021; Guvenen *et al.*, 2020; Baley *et al.*, 2022).

The rest of this article is structured as follows. We begin by introducing a new method for measuring the skill distance between occupations, based on textual data derived from job offers posted by Pôle Emploi, and showing its quantitative and qualitative validity of this method. We then present the ForCE database, and the methodological choices made when selecting the sample and variables of interest for our study. Finally, we present the results of a comparison between the career trajectories of jobseekers who have received training and those registered with Pôle Emploi who have undergone no training.

1. A New Measure of the Skill Distance Between Occupations

Our analysis of the relationship between training and the return to employment is founded upon a new measure of the distance between occupations, reflecting the differences in the skills demanded by different occupations, which can represent an obstacle to professional transitions.

There are a number of existing sources providing quality data pertaining to both occupations and professional skills. In the United States, the O*NET system provides a detailed inventory of

3. *Our checks revealed that the propensity score matching method (Rosenbaum & Rubin, 1983) yields wholly comparable results.*

skills and, for each occupation in the American classification system (the Standard Occupational Classification), an indicator for frequency of use and level of expected competence in appropriate skills. It also contains a table of related occupations (the Related Occupation Matrix) which gives, for each occupation, a ranked list of the top ten occupations to which workers could most easily transfer, in light of the similarity in the skills required. In France, the *Répertoire Opérationnel des Occupations et des Emplois* (Operational Register of Occupations and Jobs - ROME V3) occupies a comparable role: as well as defining 532 occupations and 14 sectors of activity, it contains a skills register which can be matched with the classification of occupations; for any given occupation, ROME can propose a list of possible occupations to which workers might be able to switch (the Mobilities function). The ROME register also allows us to broach the question of the skill gap between specific occupations, either by studying existing overlaps between the skills associated with each occupation, or else by looking at all of the professional trajectories implied by transitions to related occupations as per ROME's Mobilities function.

Nevertheless, using ROME V3 to study the impact of vocational training is subject to certain limitations. Chief among them is the fact that the professional transitions suggested in the database are not intended to be exhaustive: they omit to mention any number of transitions which might well be possible and pertinent, and say nothing at all about more difficult transitions. Furthermore, it splits suggested professional mobility opportunities into just two categories, whereas it would be more useful to have a sliding scale for the level of difficulty involved in moving from one occupation to another, particularly occupations which are relatively far removed from individuals' previous occupations, in terms of the skills required. Finally, the skills identified in ROME are often specific to just one or two occupations, which makes it impossible to distinguish between pairs of occupations which are slightly different and pairs of occupations which are completely different. These three shortcomings represent obstacles to quantitative analysis of the impact of training on professional mobility, since we start from the hypothesis that training would allow individuals to make career changes involving larger skill gaps.

In this section, we detail the development of a measure of skill distance between occupations, training a neural network on the text of job offers. The availability of a large corpus of textual data enabled us to produce a continuous, informative measure of this distance, even for pairs of occupations which are very different from one another. After describing these data, we proceed to detail the methodology used to construct our measure and the various validation exercises involved.

1.1. Data

The principal dataset used in this study is a body of text comprising more than 4 million job offers posted on the Pôle Emploi website between December 2018 and October 2020, along with the accompanying ROME codes. We used the text of these offers to pre-train our language model, then to train our neural network to predict an occupation code based on the text of a job offer. We then rebalanced the number of offers for each occupation with reference to the initial sample, so as not to introduce any training bias. We report descriptive statistics for the job offer dataset in Online Appendix S1 (link at the end of the article).

One of our objectives was to propose an alternative to existing measures based on the ROME V3 database. We thus used two fields from ROME to calculate alternative measures of skill distance between occupations, in order to test our own method.

- The "Mobilities" field assigns a numerical code to each possible pair of occupations: 1 if inter-occupation mobility seems possible without training, and 2 if inter-occupation mobility seems possible with minimal training to activate underlying skills. We used this field to produce an alternative measure of the distance between occupations, which we call *Graph distance*, i.e. the distance plotted on the oriented graph where the nodes correspond to the ROME codes, where the presence of an edge indicates that mobility is suggested in the Mobilities field, and the weighting corresponding to level 1 or 2 is defined in advance. We also used the Mobilities field when constructing our measure of the distance between occupations.

- The "Skills" structured field assigns a list of skills, both general and specific, to each of the 532 occupations in the ROME V3 classification. We used this field to produce an alternative measure of the distance between occupations, which we call the *Structured-field distance*, which is the cosine of the angle between the vectors representing the occupations in this space, where the component i is equal to 1 if the structured-field skill i is associated with this occupation, and 0 if not. This is analogous to our measure, with the difference that it is based on a representation in a space whose dimensions correspond to the total number of skills listed in the structured fields in ROME, established primarily on the basis of expert input.

In addition to these fields derived from the ROME classification, we incorporated information regarding the French labour market into the model's training dataset. To ensure that our measure of the distance between occupations only refers to the distance in terms of required skills – excluding other factors liable to influence the behaviour of agents in the labour market, such as gender stereotypes associated with different occupations, the social prestige attached to certain professions, or even differences in recruitment conditions between different occupations – we voluntarily restricted ourselves to data which we felt lent themselves to clear interpretation in terms of professional skills, especially the average level of education of workers in a given occupation,[4] as well as lists of pairs of initial and new occupations where we observed frequent transitions within companies, accompanied by wage increases, which we interpreted as instances of vertical professional mobility.[5]

1.2. The Neural Network

The construction of our skill distance measure was a three-part process: we began by using a language model to extract a rich semantic representation of the content of job offers, before using a neural network to extract only that content pertaining to skills, allowing us to situate occupations within a high-dimensional space, and finally we calculated the angular distance within this space, which gives us our measure of the skill distance between two occupations.

The principal methodological contribution of this paper lies in our construction of a high-dimensional spatial representation of the skills listed and required by the 532 occupations in the ROME V3 classification, which is pertinent to the quantitative analysis of professional mobility, while also allowing for geometric interpretation. The measure of the skill distance between occupations used in the rest of this article is a direct by-product of this construction. We use FlauBERT (Le *et al.,* 2020), a pre-trained French language model, to analyse the text of job offers.

Our decision to use a neural network to generate our representation is what sets this work apart from the existing literature on the multi-dimensionality of skills. Previous studies have predominantly relied upon methods designed to construct indices or reduce dimensions (principal component analysis, correspondence analysis, etc.), which are mathematically simple to define but yield representations whose pertinence to skills analysis is far from certain – particularly for the purposes of quantitative analysis. We opted for an alternative approach, a supervised method in which all of the pretext tasks used to train the model can be easily explained, are directly connected to the purpose of the representation, and yield a geometrically coherent result. This allows for easier interpretation of our representation, although it does make the construction process more complex.

The fact that our construction of a spatial representation of skills is based on textual data opens the door to certain risks. It is possible that the text of the job offers contains elements unrelated to skills, which are of little use when predicting the associated occupation. The style of writing, the name of the company and the location of the job may all have an influence on our representation. We might also find that, depending on the occupation, job offers tend to demand more skills than the post truly requires, or different skills during different phases of the labour market cycle, as observed by Deming & Kahn (2018). Nevertheless, the size of our corpus of job offers, and the shallow depth of the neural network, allow us to be optimistic that the network will be capable of focusing on those features of the text with the greatest predictive capacity, cutting through the noise to get to the skills descriptions.

1.2.1. Objectives

We thus used a neural network to situate the 532 occupations listed in the ROME classification within a 20-dimensional space, so that the relative positioning of two occupations would provide information as to the likelihood of professional transition from one to the other. The dimensions of this space are not intended to be considered in isolation as measurements of the importance of a specific (or even clearly interpretable) skill for a given occupation. However, the vector formed by these 20 dimensions contains a composite representation of the

4. To ascertain this value we use the highest level of qualification listed in the Pôle Emploi records for those registered in 2018. This information is derived from the ForCE matching.
5. These vertical transitions are constructed using administrative data for the year 2019 (All employees database – job position data - Base tous salariés, fichier "Postes"*, INSEE).*

skills required by a given occupation, and this vector lends itself to interpretation. Once the occupations have been situated in the space, we use the cosine distance between vectors to measure distances between occupations in terms of skills.

Expressed in formal terms, the representation R is the function which associates each occupation in the ROME classification with its corresponding vector in this space.

$$R: \begin{array}{ccc} \{\text{Codes ROME}\} & \rightarrow & \mathbb{R}^{20} \\ x & \mapsto & R(x) \end{array}$$

The representation is constructed in such a manner that the geometry of this space can be clearly interpreted as it pertains to skills and professional mobility.

(i) The angle between vectors $R(x)$ and $R(y)$ must reflect the extent to which the skills required by occupations x and y are similar. The cosine of this angle is our measure of skill distance between occupations.

(ii) The norm $\| R(x) \|$ must reflect the expected degree of expertise in the skills required by occupation x.

(iii) The projection of the vector u on the line spanned by $R(x)$ must reflect the degree of expertise in the skills required by occupation x possessed by a person whose skills are represented by u.

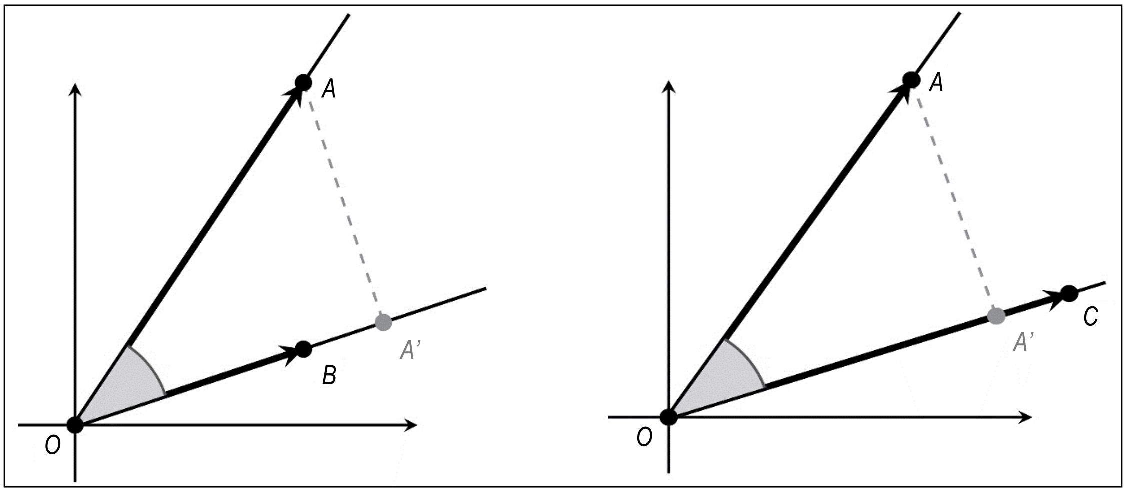
(iv) All components of the vector $R(x)$ are positive and represent dimensions of the skills required by occupation x.

The distinction between required skills (direction of the vector) and the degree of expertise in these skills (norm of the vector) is inspired by the questions contained in several of the existing data sets linking skills and occupations, such as the American O*NET classification. In practice, we use the direction of the vectors in the representation of occupations (see below), but their norm also plays a role in the process of constructing that representation.

Figure I illustrates the geometric interpretation of our representation. One consequence of the third property, relating to skills transfer during professional transition, is that the dimension of the space implies that most occupations will be far from each other in terms of the skills they require. In other words, moving from one of these occupations to another would involve starting from zero. We opted to work in 20 dimensions, which means that our representation must cover the skills shared among the 532 occupations in the ROME classification, while also leaving enough flexibility for these skills to be represented in a low-dimensional model.

Our proposed geometric frame has two properties which may appear counter-intuitive. On the one hand, the representation introduces a form of imperfect substitutability between skills. Possessing a very high level of expertise in a given field guarantees a minimum level of ability in every other field which has at least one skill in common. In this respect, the representation does not do a good job of handling situations where transition is totally impossible. Moreover,

Figure I – **Geometric interpretation of the representation *R* of occupations**



Note: Representation of the transition from an initial occupation A to a new occupation B within the skill space. Our measure of distance between occupations, which reflects the skill gap between occupations A and B, corresponds to the cosine of the angle in grey. Occupation A requires a higher level of skills than Occupation B, hence the distance between the two is greater. If an individual whose skills correspond exactly to those demanded by Occupation A were to transition to Occupation B, his degree of expertise in the skills required by B would be the distance from the origin O to point A . In this case, the degree of expertise is sufficient to make the transition from A to B possible. However, this same individual would not be able to move to Occupation C , which involves the same skills as Occupation B but demands a higher level of expertise.

the binary relation between occupations which indicates that professional transition is possible is not usually transitive; in other words, it may be possible to move from occupation A to occupation B, and from occupation B to occupation C, but never directly from A to C. The underlying idea is that this binary relation reflects the possibility of undertaking professional mobility within a reasonable time frame without receiving specific training. But the sum of two "reasonable" periods of time may no longer be so reasonable.

This geometric representation of the feasibility of professional transitions within the skills space was not used directly in the construction of our measure. Nonetheless, it does illustrate the partial reuse of existing skills when undertaking professional mobility. The further away we move from the initial occupation in terms of the skills required, i.e. the bigger the angle, the smaller the norm of the projection will be, i.e. the less useful existing skills will be. So, for a given level of expertise in the skills required by a new occupation, it is more feasible to make the transition to an occupation with a small angular distance. This highlights the advantage of using angular distance as a measure of the difficulty associated with a professional transition.

In addition to this representation of occupations, our neural network also allows us to assign every job offer to a position within this 20-dimensional space. We use F to represent the corresponding function.

$$\begin{array}{rccl} F: & \{\text{job posting text}\} & \rightarrow & \mathbb{R}^{20} \\ & x & \mapsto & F(x) \end{array}$$

This representation of the texts of our job offer corpus has two advantages:

- It can be used to learn the representation R of occupations. The primary task when training the neural network consisted of comparing the text t of a job offer with an occupation x, by means of their representatives $F(t)$ and $R(x)$, in order to predict the ROME code of the occupation targeted by the job offer.

- Once training was complete, we used the network to conduct a qualitative analysis of the resulting representation. It is also possible to modify the text t of a job offer at the input stage, then observe how the changes affect the representative $F(t)$.

1.2.2. Architecture of the Neural Network

The neural network comprises three blocks, as illustrated in Figure II.

Block 1 is a language model. It takes a textual input and generates a representation of this text in 768 dimensions, encompassing a broad array of semantic problems. We used a pre-trained version of the model, with parameters which remained fixed while training the other Blocks.

Block 2 is a neural network with three layers. It takes as its input the representation of a job offer produced by Block 1, then produces a representation of this offer in 20 dimensions. This step dispenses with the 768 initial dimensions, retaining only the information pertaining to skills and professional mobility. Combined with Block 1, it forms a function F which assigns the text of a job offer to its 20-dimensional representation.

Block 3 is an additional layer which stores the representation R of the 532 occupations in the ROME classification, in 20 dimensions. Comparing the output of Block 2 with the output of Block 3 allows us to assess the performance of the network in its various tasks, and to adjust the training parameters accordingly.

The parameters of Block 1 (dotted background in Figure II) were retrieved from a pre-trained model (FlauBERT. Le *et al.,* 2020), with some slight re-training on a semi-supervised basis, focusing on our corpus of job offers independently of the rest of the model, before freezing these parameters during training of Blocks 2 and 3. The parameters of Blocks 2 and 3 (dark grey layers in Figure II) were random to begin with, then trained on the text of the job offers as explained in the next sub-section.

1.2.3. Training Tasks and Associated Penalties

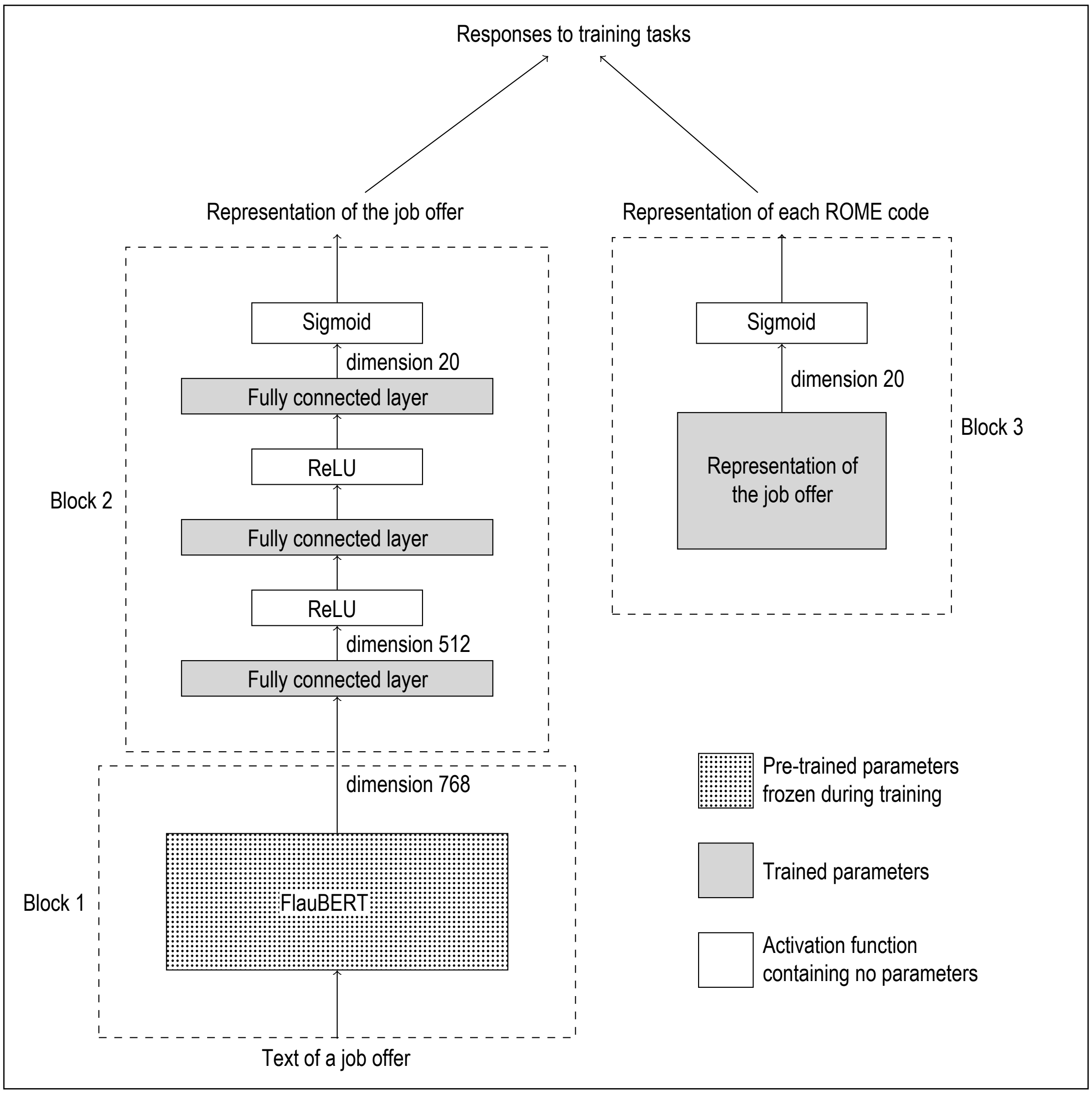
The representation of occupations R is one of the parameters of the neural network, which evolve as the neural network is trained to perform certain tasks. Choosing appropriate training tasks and associated penalties allows us to impose upon the representation the geometric properties detailed above.

We used the following task-loss pairings to train our neural network.

Predicting the ROME Code Associated With a Job Offer - *WARP Loss*

We decided to adapt the *Weighted Approximate-Rank Pairwise* loss (or WARP) proposed by Weston *et al.* (2011) to the context of our problem. For each text t representing a job offer, we calculated the angle of its representative $F(t)$ with the representative $R(x)$ for each occupation

Figure II – **Diagram of our neural network**



x in the ROME classification. We expected the angle between $F(t)$ and the representative $R(x_0)$ for the occupation corresponding to the job offer to be very small. We thus penalised the network for each occupation x so that the angle between $F(t)$ and $R(x)$ remained smaller than the angle between $F(t)$ and $R(x_0)$.

$$l_{\text{WARP}} = \sum_{\text{ROME codes } x \neq x_0} \left| F(\mathrm{t}) \cdot \left(R(x) - R(x_0) \right) \right|_+$$

where $|\cdot|_+ = max(0,\cdot)$.

Predicting the Mobility Options Suggested in the ROME Classification - Triplet Loss

Triplet loss is a concept whose origins lie in image recognition. If we take three photographs of human faces, with the first two showing the same individual and the third featuring a different person, a facial recognition algorithm should find that the first image bears closer resemblance to the second than to the third.

Similarly, we created triplets of occupations, with an initial occupation x_0, an occupation x which is suggested in the Mobilities section of the ROME classification as a possible career move for individuals in y, and a third occupation which is not suggested as a possible destination for professional transition. We then picked, at random, a job offer corresponding to each of these occupations.

We would expect mobility from x_0 to x to be a better option than moving from x_0 to y. The transition needs to be feasible in terms of the skills required, without too much of a decline in the level of expertise compared with the initial occupation. The following equation can thus be used to determine the extent to which a transition from one occupation to another would be "recommended":

$$d\left(x_0,x\right)=max\left(1-\frac{\|R\left(x\right)\|}{\|R\left(x_0\right)\|},1-\frac{R\left(x_0\right)\cdot R\left(x\right)}{\|R\left(x\right)\|^2}\right)$$

Triplet loss can thus be defined as the difference, with reference to this equation, between the pairs $\left(x_0,x\right)$ and $\left(x_0,y\right)$.

$$l_{Triplet}=|\,d\left(x_0,x\right)-d\left(x_0,y\right)|_+$$

Predicting the Expected Level of Expertise in the Skills Required by an Occupation x - Norm Loss

For each occupation x, we compared the norm of representative $R\left(x\right)$ to a value e_x (normalised between 0 and 1) representing the average qualification level of workers in this occupation.[6]

$$l_{\text{Norm}}=\sum_{\text{ROME codes } x}\left(\|R\left(x\right)\|-e_x\right)^2$$

Predicting Level of Expertise - Vertical Loss

In many sectors, vertical mobility requires candidates to have acquired the skills associated with lower ranks of the same profession. We would thus expect that, for initial-new occupation pairings which are often associated with vertical mobility in the labour market data, the initial occupation would demand fewer skills than the new occupation, in all dimensions encompassed by the notion of skills.

Using $x\prec y$ to denote the set of occupations y for which transitioning from occupation x constitutes a vertical mobility, we define *vertical loss* as

$$l_{\text{Vertical}}=\sum_{x\prec y}\sum_{i=1}^{20}|\,R(x)_i-R(y)_i\,|_+$$

The various penalties associated with these tasks are normalised, passed through the logarithm and added up to calculate a total loss l, which gives us:

$$l=\log\left(1+l_{\text{WARP}}\right)+\log\left(1+l_{\text{Triplet}}\right)+\log\left(1+l_{\text{Norm}}\right)$$
$$+\log\left(1+l_{\text{Vertical}}\right)$$

1.3. Results

We qualitatively and quantitatively verified the skill distance measures obtained using our spatial representation of occupations. In the appendix to this article, we provide some initial indications of the performance of the algorithm in the pretext tasks, as well as a qualitative validation exercise.

When it came to quantitatively validating our measure, our principal approach consisted of evaluating its capacity to predict professional transitions actually recorded in the labour market.

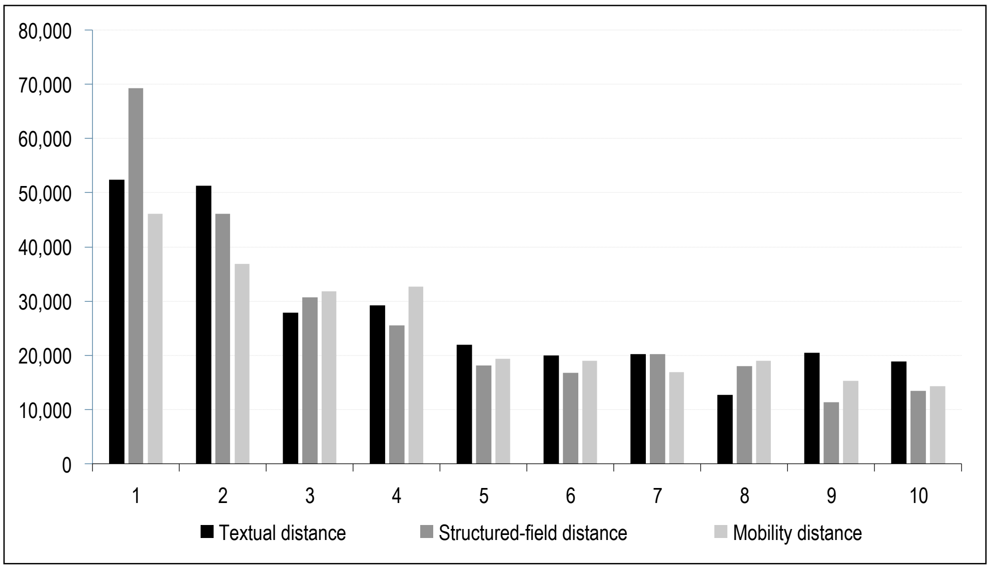
Our measure is not designed to perfectly predict observed transitions, as career changes may be influenced by any number of factors which have nothing to do with professional skills, such as personal aspirations, gender stereotypes or conditions on the local labour market. Nonetheless, it seems reasonable to assume that there would be more professional transitions between occupations which are relatively close to one another, in terms of the skills involved.

We compared the predictive power of our own measure of distance between occupations with two alternative measures: first, the distance on the graph of the professional mobilities suggested by the ROME classification, and second the cosine distance between the skills vectors associated with initial and new occupations in the ROME classification.

Figure III represents, for each of the three measures of distance between occupations and for each level k (between 1 and 10), the number of professional transitions observed from an initial occupation towards the closest k-th occupation. This enables us to assess the capacity of each measure to identify the most likely transitions, and to observe the quality of the measures for more distant occupations. We can thus see that our measure of distance (noted Textual distance in the graph) performs better than the distance constructed using the suggestions from the ROME classification (Mobility distance) at 7 of the 10 levels. Furthermore, adding up the transitions explained by the k closest target occupations, our measure is superior for all k values from 1 through 10. As for the measure based on the skills defined in the ROME classification V3 (Distance in skills), we can see that it has impressive predictive power for the closest occupations, but is poor at predicting more remote occupations. This is not hard to understand, because the definitions of skills used in the ROME classification V3 are narrow and do not allow for easy generalisation (a shortcoming which has been taken into consideration in the ongoing work to create version 4.0 of the ROME classification). Our measure, however, which was constructed using diverse and unstructured textual data, seems more capable of identifying broad skills categories and making generalisations, as it outperforms its rivals at the highest deciles.

6. In practice, this level is constructed as the average level of education observed in the occupation, normalised between 0 and 1.

Figure III – **Number of transitions explained by measures of distance between occupations**



Note: This bar chart represents the number of transitions observed in the labour market in 2019, towards the 10 occupations closest to the initial occupation according to different measures of distance between occupations. We thus compare the textual distance that we constructed to the distances constructed using the Mobilities and Skills structured fields from ROME V3. This diagram makes no distinction between transitions made with or without some form of training.
Source: *Base tous salariés, fichier "Postes"*, 2019, INSEE; authors' calculations.

Table 1 contains the results of our regression analysis, for all pairs of occupations between which at least one transition was observed, of the logarithm for the number of recorded transitions and the different normalised measures of the distance between those occupations. This allows us to compare the capacity of the different measures to predict long distance transitions, i.e. between occupations which are quite distant in terms of the skills they demand, compared with the previous diagram which focused on the most likely transitions. Our measure of distance between occupations does a better job of accounting for variance than the two other measures, and the correlation coefficient is higher. It should come as no surprise that measures of the distance between occupations which focus primarily on skills are only capable of explaining a small proportion of the transitions observed, since they do not take into consideration any of the numerous other factors which may influence professional mobility.

2. Vocational Training, Returning to Work and Career Trajectories

In this section we consider the career progress of jobseekers who undertake training, in terms of their return to employment and their career trajectories in general, comparing their experiences with those of jobseekers who have not undertaken training but are comparable with regard to other observable properties. We begin

Table 1 – **Explanatory power of measures of distance between occupations**

	log(transitions)	log(transitions)	log(transitions)	log(transitions)
Mobility distance	−0.409 (0.0052)			−0.288 (0.0049)
Structured-field distance		−0.4501 (0.0051)		−0.287 (0.0049)
Textual distance			−0.609 (0.0052)	−0.495 (0.0052)
N	149,209	149,209	149,209	149,209
R^2	0.0409	0.0506	0.0935	0.1388

Note: This table contains the results of four regressions for which the dependent variable was the logarithm for the number of transitions between occupations. The first two explanatory variables are distance measures constructed using the Mobilities and Skills structured fields from ROME V3, which we compare with the textual distance constructed for the purpose of this study. Each measure of distance between occupations has been normalised so that the coefficient directly expresses the correlation between the explained and explanatory variables.
Source: *Base tous salariés, fichier "Postes", 2019*, INSEE; authors' calculations.

by detailing the administrative data sources we used and their preliminary processing, then offer a brief explanation of our chosen empirical strategy, before presenting our results.

2.1. Data and Structure of the Sample

We used data derived from the Training, Unemployment and Employment (ForCE)[7] programme, which merges:

- The Historical File on jobseekers (FH, *Fichier historique des demandeurs d'emploi*), containing information on all jobseekers registered with Pôle Emploi in the 10 years preceding the latest edition of ForCE;

- The regional database of participants enrolled in vocational training (Brest, *base régionalisée des stagiaires de la formation professionnelle*), containing details of training courses completed since 2017 by all jobseekers who have at one point undertaken vocational training;

- The manpower movement database (MMO – *Mouvements de main-d'œuvre*, based on data derived from the nominative social declaration, or DSN – *Déclaration sociale nominative*), containing information on employment contracts involving all employees in the private sector since 2017,[8] for all jobseekers present in the HF of the latest edition of ForCE.

We constructed our analytical database using the same broad principles employed by Chabaud *et al.* (2022) in their study of the impact of vocational training on getting jobseekers back into employment, using similar data. We began by creating a database for each month between January 2018 and December 2020, containing all jobseekers registered during that month m (excluding category E / administrative category 5),[9] after first combining any periods of registered unemployment less than 30 days apart. In order to focus our analysis on jobseekers undertaking training for the first time, we used Table P2 in the Historical File to exclude any jobseekers enrolled in training courses in 2017. The sample was matched with the Brest database in order to identify all jobseekers enrolling on a training course for the first time in month m, the group we were interested in for that month – while also excluding the control group of individuals who had received training before month m. We also excluded training programmes regarded as being directly connected to recruitment schemes (POEI, POEC, AFPR), which did not fall within the scope of our study.[10] Finally, matching with the MMO database enabled us to retrieve information regarding the contracts held by jobseekers *(i)* before month m and *(ii)* for each of the 24 months following month m. In so far as this study is devoted to the career trajectories of jobseekers, our principal sample restricts the population of jobseekers to people who have held a stable job at some time in the 12 months preceding month m. We thus identified a reference occupation for each jobseeker which, when linked with the characteristics of the job they found after month m, enabled us to construct the principal variables of interest for this study (skill distance between the two occupations, differential in recruitment difficulties, nature of employment contract, etc.).

We thus retained a large amount of information regarding individuals present in the HF, including their age, gender, marital situation, number of children, level of education and qualification, disability status, residence in urban/rural areas, willingness to undertake geographic mobility, registered unemployed status, reason for registration, nationality, desired occupation, reservation wage, time in unemployment as of month m, total time in unemployment and number of periods of unemployment before this period. These control variables allow us to correct for observable differences between jobseekers who have undergone vocational training and those who have not.

2.2. Empirical Strategy

Let us begin by defining a variable $D_i \in \{0,1\}$ which tells us whether an individual i started training in month m $(D_i = 1)$ or not $(D_i = 0)$. For a variable of interest Y_i (for example, return to employment within 24 months of the month m in which training began), we can define two potential values (using the standard terminology of causal inference) indicating the value of Y_i for an individual i who either did not begin training in the month i (m) or did indeed begin training $(Y_i(0))$. Intuitively, the conditional independence assumption supposes that a treated individual $(D_i = 1)$ would have, in the event that they did not enter training, a similar fate to a control individual i' $(D_{i'} = 0)$ with observable properties X similar to those of i. In formal terms, this can be expressed as follows:

7. For the year 2023T2.
8. Since 2022, information on the current contracts of a large proportion of public sector employees has also been available.
9. Category E / Administrative Category 5 contains all registered jobseekers currently in full time employment, and who are therefore under no obligation to find new employment.
10. The Brest database allows us to distinguish Individual Operational Preparations for Employment (POEI – Préparations Opérationnelles à l'Emploi Individuelles), Collective Operational Preparations for Employment (POEC – Préparations Opérationnelles à l'Emploi Collectives), and Pre-Recruitment Training Actions (AFPR – Actions de Formation Préalables au Recrutement) from the Individual Training Subsidies (AIF – Aides Individuelles de Formation) on which we are focusing.

$$Y_i(0) \perp D_i \mid X.$$

This hypothesis, if verified, enables us to identify the average treatment effect on the treated (ATT), i.e. the average impact of beginning training in month m on the individuals concerned. When interpreting these results, it is important to bear in mind that the treatment we are interested in is the fact of entering training for the first time *during month* m, in comparison with all other scenarios (including no training at all, *or* entering training after month m).

However, it is worth noting that in this case it is probable that those jobseekers choosing to enrol in training are different from untrained individuals in terms of unobservable characteristics (motivation, career ambitions, etc.) likely to be significantly correlated with return to employment and a positive career trajectory in general. This form of bias – which impedes causal interpretation of analyses based on the conditional independence assumption which do not control for differences in observable characteristics – has been extensively discussed in the literature (Lalonde, 1986). It thus seems more prudent to interpret our results as measurements of the existing correlation between vocational training and return to employment, corrected for differences in observable characteristics.

With regard to estimation, there are various ways to control for the differences associated with observable factors X, especially when the dimensionality of X is high. The first is propensity score matching, a method which has been widely used to solve the dimensionality of X problem since it was first proposed by Rosenbaum & Rubin (1983). Our preferred method, known as Double Debiased Machine Learning (DML) (Chernozhukov *et al.*, 2018), relies on the use of nonparametric estimators for the conditional expectation of the variable of interest and the propensity score, which can then be combined to create an estimator capable of withstanding erroneous specification of one of the two terms.[11] Although this solution yields better statistical properties than the classic propensity score matching method, the question of whether or not the DML estimator is biased is just as dependent on the validity of the conditional independence assumption.

2.3. Characteristics of our Sample

In this sub-section we describe the sample of jobseekers used in our analyses. As noted above, the population we are interested in is limited to individuals for whom we can identify a reference occupation and who, as a result, are relatively close to the world of work. As mentioned previously, we also excluded all jobseekers undertaking training directly linked to recruitment programmes (POEI, POEC and AFPR).

Table 2 compares jobseekers known to have held a stable employment contract (permanent post, or fixed-term contract of more than 6 months) within the previous 12 months (the population we set out to study), with those for whom this was not the case. The first sub-population is the only group for which we are able to study the link between training and skill acquisition, as well as the differential in recruitment tension between the initial and subsequent occupations. This sub-population appears to be younger, with a smaller proportion of women and more university graduates; people in this group have been unemployed for less time, and are more likely to undertake training.

Table 3 presents the characteristics of the jobseekers who make up the population of interest to our study, depending on the type of training they have undertaken. We make a distinction between jobseekers not undertaking training and jobseekers taking up training options of 30+ hours, while also considering the type (leading to a diploma, or not) and duration (more or less than 420 hours, equivalent to 3 months of full-time study) of such training programmes.[12]

Compared with those jobseekers who have not enrolled in 30-hour training programmes, the sub-population of jobseekers embarking upon their first 30-hour (or more) training unit are, on average, younger, have been unemployed for less time, and comprise a higher proportion of women and university graduates. These observations are even more salient for those sub-populations of individuals enrolling on courses leading to diplomas, or long training courses involving more than 420 hours of teaching. One more surprising observation is the absence of any notable difference between these populations in terms of career change aspirations. We measured these aspirations by considering the proportion of jobseekers reporting that they were seeking a different occupation from their last stable

11. In terms of concrete implementation of this method, packages in Python and R (the language used for this study) are available at: https://docs.doubleml.org/stable/index.html.

12. The number of jobseekers beginning training is subject to major seasonal variation. There is a clear peak in September. We decided to use statistics for individuals enrolling on training programmes between January 2018 and December 2020.

Table 2 – **Descriptive statistics for jobseekers who had (or did not have) stable jobs within the 12 months preceding the month of the study**

	No stable job in the preceding year	With stable job in the preceding year
Female (%)	53.4	48.9
Age (%)		
Under 25	10.2	12.4
Age 25-50	60.1	64.5
Over 50	25.1	18
Level of education (%)		
No high school diploma	50.6	43.2
High school diploma (baccalaureate)	22.1	22.6
Higher than baccalaureate	27.3	34.3
Time in unemployment (months)	31.7	12.6
Training (%)	5.2	8.8
Diploma training (%)	2.1	4.0
Training > 420 hours (%)	2.4	3.7
Observations (thousands)	4,118	908

Source and field: ForCE data, DARES. All jobseekers registered between January 2018 and December 2020, excluding all jobseekers undertaking training directly linked to recruitment programmes (POEI, POEC and AFPR).

Table 3 – **Descriptive statistics by training status for the population of jobseekers who had been in stable employment within the previous year**

	No training	Training	Diploma training	Training > 420 hours
Female (%)	48.9	49.8	51.3	55.7
Age (%)				
Under 25	12.4	12.4	13.3	18.6
Age 25-50	64.4	69.3	71.6	69.3
Over 50	18.1	14.0	10.7	7.7
Level of education (%)				
No high school diploma	43.3	38.5	33.7	32.7
High school diploma (baccalaureate)	22.5	26.5	29.2	31.9
Higher than baccalaureate	34.2	35.0	37.1	35.4
Time in unemployment (months)	12.7	8.3	8.3	8.4
Intended change of occupation (%)	67.5	69.2	69.9	70.8
Distance (when change occurred)	0.51	0.53	0.53	0.54
Observations (thousands)	900	8.0	3.7	3.4

Source and field: ForCE data, DARES. All jobseekers registered between January 2018 and December 2020 after termination of an employment contract (permanent post or fixed-term post of more than 6 months) within the preceding 12 months, excluding all jobseekers undertaking training directly linked to recruitment programmes (POEI, POEC and AFPR).

post.[13] Almost three quarters of the individuals surveyed reported that they were looking for a different occupation from their last post – a proportion which varies little if at all, regardless of treatment status. Moreover, the scale of the skill changes involved – as measured by the skill gap between their previous occupation and the occupation they wish to take up – is comparable for all individuals, whether or not they were beginning training, and irrespective of the nature of that training.

13. For the purposes of this exercise, we used the occupations found in the most detailed version of the FAP classification, containing 225 posts. Using less detailed data, for example FAP 87, does not change the result. Using the FAP classification allowed us to bridge the gap between the ROME classification used in the HF and the PCS system used in the MMOs.

Our identification and estimation strategies were based on the correction of observable differences between control and treated jobseekers, using the DML method (see above). In practice, the control variables used by our algorithm included level of qualification, age, gender, type of contract sought, level of training, experience on the labour market, whether or not jobseekers live in priority urban development zones, marital status, administrative category of Pôle Emploi registration, nationality, the reason for their most recent registration with Pôle Emploi, whether or not they are under obligation to seek employment, the level of their declared reservation wage, their mobility preferences, their desired occupation, the employment zone associated with their place of residence, the number of discrete periods of unemployment and the total time they have been registered with Pôle Emploi over the past ten years.

2.4. Results

All of the results presented in this section were obtained using the *Double Debiased Machine Learning* technique (see above), and are broadly comparable to those obtained using a more familiar propensity score matching method.

Table 4 summarises our results on the correlation between vocational training and return to employment within different time frames (3, 6, 12 and 24 months after beginning training), for all jobseekers enrolling on training courses for the first time between January 2018 and June 2020, and having previously held a stable job in the twelve months before they began training. We also make a distinction between return to employment after training courses leading to diplomas and return to employment in general, before adding a further restriction to include only those jobseekers returning to work in stable jobs (permanent contracts or fixed-term contracts of more than 6 months). As has been extensively documented in the literature (Card *et al.*, 2018), we observed a "lock-in" phenomenon, which is to say a negative correlation between enrolling on a training programme and returning to employment in the short term, due to the time required to complete the training. The correlation between training and return to employment is positive 12 months after beginning training, and remains positive while becoming more pronounced 24 months after beginning training. Our results, which focus on a sub-population relatively close to finding employment, are both qualitatively and quantitatively different from the results reported by Chabaud *et al.* (2022). From a qualitative standpoint, the lock-in effect appears to be more evident and more lasting for our chosen population than it is for the entirety of the population registered with Pôle Emploi. From a quantitative perspective, the corrected differential between trained and untrained jobseekers after 24 months (for our target population) is around 20% less than the result obtained by Chabaud *et al.* (2022) in a study encompassing the entirety of the population registered with Pôle Emploi. These differences may arise from the fact that the individuals in our sample are, by construction, closer to returning to employment than the average jobseeker. It may also be related to the type of training programmes we included in our model. As such, enrolling on a training programme is more likely to significantly reduce return to employment opportunities for this sub-population, for whom employment opportunities are plentiful, even without undertaking further training.

Table 4 shows that while the initial lock-in effect is stronger for training programmes leading to diploma qualifications (which are often long courses of education), the differential between graduates/non-graduates of such programmes after 24 months is greater than the differential for training programmes as a whole. We find something broadly similar when we look at return to stable employment (permanent contracts or fixed-term contracts of more than 6 months). However, the differential between graduates of diploma courses and non-graduates after 24 months is not substantially different from the figure for training courses as a whole. The lock-in effect, however, appears to be more long-lasting: still significant 12 months after training.

Table 5 summarises our principal results on the relationship between training and the career trajectories of jobseekers undertaking various types of training courses (all training programmes and programmes leading to diplomas) 24 months after beginning that training. As explained above, our results concern a sample of jobseekers who had previously been in stable employment within the preceding twelve months. In this table, the dependent variable for return to employment within 24 months is broken down according to the distance between the original and subsequent occupations. We thus distinguish between return to employment in the original occupation ($d = 1$), in an occupation very close to the original occupation ($d \in [2;5]$), in an occupation close to the original occupation ($d \in [6;20]$), and finally in an occupation requiring very different skills from those associated with the original occupation ($d > 20$). We can thus observe that

Table 4 – **Observed differences in return to employment between trained and untrained individuals**

	(1) 3 months	(2) 6 months	(3) 12 months	(4) 24 months
		All types of employment		
All types of training	−0.100 (0.002)	−0.089 (0.002)	0.007 (0.002)	0.069 (0.002)
Diploma training	−0.119 (0.002)	−0.119 (0.003)	0.018 (0.003)	0.086 (0.003)
		Stable employment		
All types of training	−0.063 (0.001)	0.06 (0.002)	−0.011 (0.002)	0.033 (0.002)
Diploma training	−0.078 (0.002)	−0.082 (0.002)	−0.013 (0.003)	0.04 (0.003)

Note: This table contains the results of separate regressions for 4 different dependent variables (in columns, corresponding to return to employment within different time frames) with 2 explanatory variables (the rows, corresponding to the different types of training). The upper section of the table corresponds to return to all types of employment within different time frames, while the lower section repeats these analyses but uses return to stable employment as the only dependent variable (permanent post or fixed-term contract of more than 6 months). Standard errors clustered by occupation × employment zone in parentheses.

vocational training reduces the probability that a jobseeker will return to employment in their original occupation, has a virtually neutral effect on similar occupations, and substantially increases the probability of returning to employment in an occupation requiring different skills from their original occupation. These effects are even more evident when we focus exclusively on training courses leading to diplomas. We can thus say that training in general, and diploma programmes in particular, do appear to coincide with a reallocation of labour supply towards occupations requiring different skills than the posts held before training. This result is striking in so far as it suggests that the entire correlation between return to employment and vocational training involves professional transitions towards very different occupations.

Is the boost in employability which we observed for participants in vocational training backed up by a greater likelihood of moving to a sector of the economy which is looking to recruit? In an attempt to answer that question, Table 6 breaks down the relationship between training and return to employment, looking at whether or not recruitment difficulties are more acute in the market sectors to which trained jobseekers move. The data used to calculate this breakdown are obtained by means of a regression analysis of the return to employment of untrained jobseekers and fixed market effects (employment area × FAP), controlling for the individual characteristics of the jobseekers present in each market. This provides us with an indicator (rate of return to employment for each market) which avoids the familiar measurement problems associated with the non-observable dimensions of the efforts made by jobseekers and businesses.[14] Table 6 shows that, for diploma and non-diploma training programmes alike, the return to employment differential after 24 months does not seem to be driven by more jobseekers switching to

14. DARES use a comparable measurement (labour dispersal by market) to construct a synthetic indicator of tensions on the labour market.

Table 5 – **Differences in return to employment after 24 months for trained and untrained individuals, as a function of the skill distance between their original and new occupations**

	(1) Initial occupation $(d = 1)$	(2) Very similar occupation $(d \in [2;5])$	(3) Similar occupation $(d \in [6;20])$	(4) Different occupation $(d > 20)$
All types of training	−0.044 (0.001)	0.005 (0.001)	0.006 (0.001)	0.08 (0.002)
Diploma training	−0.052 (0.002)	0.009 (0.002)	0.005 (0.002)	0.095 (0.003)

Note: This table contains the results of separate regressions for 4 different dependent variables (the columns) with 2 explanatory variables (the rows, corresponding to the different types of training). For example, the coefficients of column (2), Very similar occupation (d∈[2;5]), correspond to the training/no training differential in return to employment (after 24 months) in one of the 4 occupations regarded as being closest to the individual's previous occupation, according to our measure of distance between occupations. Standard errors clustered by occupation × employment zone in parentheses.

Table 6 – **Differences in return to employment after 24 months for trained and untrained individuals, with reference to recruitment conditions in the new occupation**

	(1) Recruitment difficulties less acute than in initial occupation	(2) Recruitment difficulties greater than in initial occupation
All types of training	0.046 (0.002)	0.045 (0.002)
Diploma training	0.050 (0.003)	0.059 (0.003)

Note: This table repeats the analyses from Table 4 but specifies, when constructing the dependent variables, whether individuals returned to unemployment in market sectors where recruitment difficulties were more (or less) acute than in their previous occupations. Standard errors clustered by occupation × employment zone in parentheses.

markets where recruitment shortages are more common than they were in their original occupations. All in all, the impact of training (whether or not it leads to a diploma) on the probability of transitioning towards an occupation where the recruitment demand is stronger appears to be comparable to the probability of transitioning towards an occupation where the recruitment demand is actually weaker. If the aim of the training system is to redress imbalances of supply and demand between different labour markets, this result may appear to be somewhat disappointing. It suggests that refocusing the range of training courses on offer would serve to redirect labour supply towards those sectors experiencing recruitment difficulties; as things currently stand, the impact of training on such transitions appears to be broadly neutral.

* *
*

Does vocational training serve to mitigate structural imbalances in the labour market? This study seeks to provide some form of answer to this question, comparing the career trajectories of jobseekers with and without vocational training. To this end, we constructed an original measure of the skill gap between different occupations, using the text of job offers posted by Pôle Emploi. We used this measure to study the professional transitions undertaken by jobseekers with or without training. For a sample of jobseekers relatively close to finding employment, our results show that, compared with jobseekers without vocational training, trained jobseekers tend to make professional transitions over greater distances, in terms of the skills involved. Our results depend on a conditional independence hypothesis – which is strong in this context – regarding training decisions, and must thus be interpreted with a degree of prudence. In terms of reallocative capacity, the increased likelihood of return to employment after vocational training does not appear to be driven by a surplus of jobseekers switching to occupations where recruitment difficulties are more acute. This result suggests that greater effort to ensure that the range of vocational training options on offer are systematically focused on the skills demanded by occupations in need of manpower would serve to boost the reallocative impact of vocational training. ☐

Link to the Online Appendix:
www.insee.fr/en/statistiques/fichier/8679062/ES547_Frick-et-al_Online-Appendix.pdf

BIBLIOGRAPHY

Baley, I., Figueiredo, A. & Ulbricht, R. (2022). Mismatch Cycles. *Journal of Political Economy*, 130(6). https://doi.org/10.1086/720461

Bana, S. H., Brynjolfsson, E., Rock, D. & Steffen, S. (2019). Job2Vec: Job Title Benchmarking with Collective Multi-View Representation Learning. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). Association for Computing Machinery, New York, NY, USA, 2763–2771. https://doi.org/10.1145/3357384.3357825

Barnichon, R. & Figura, A. (2015). Labor Market Heterogeneity and the Aggregate Matching Function. *American Economic Journal: Macroeconomics*, 7(4), 222–249. https://doi.org/10.1257/mac.20140116

Basco, S., Liégey, M., Mestieri, M. & Smagghue, G. (2024). The effect of import competition across occupations. *Journal of International Economics*, 153, 104001. https://doi.org/10.1016/j.jinteco.2024.104001

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis with special reference to education*. New York & London: Columbia University Press for the National Bureau of Economic Research. https://www.nber.org/books-and-chapters/human-capital-theoretical-and-empirical-analysis-special-reference-education-first-edition

Card, D., Kluve, J. & Weber, A. (2018). What Works? A Meta Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association*, 16(3), 894–931. https://doi.org/10.1093/jeea/jvx028

Chabaud, M., Bucher, A., Givord, P. & Louvet, A. (2022). Quelles sont les chances de retour à l'emploi après une formation ? DARES, *Document d'études* N° 261. https://dares.travail-emploi.gouv.fr/publication/quelles-sont-les-chances-de-retour-lemploi-apres-une-formation

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Dawson, N., Williams, M. A. & Rizoiu, M. A. (2021). Skill-driven recommendations for job transition pathways. *PLoS ONE*, 16(8), e0254722. https://doi.org/10.1371/journal.pone.0254722

Deming, D. & Kahn, L. B. (2018). Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. *Journal of Labor Economics*, 36(S1), S337–S369. https://doi.org/10.1086/694106

Eckardt, D. (2022). Are Chemists Good Bankers? Returns to the Match between Training and Occupation. *Working Paper*. https://conference.iza.org/conference_files/TAM_2022/eckardt_d29661.pdf

Fontaine, F. & Rathelot, R. (2022). Le marché du travail français à l'épreuve de la crise sanitaire. *Notes du Conseil d'analyse économique*, 71(2), 1–12. https://www.cairn.info/revue-notes-du-conseil-d-analyse-economique-2022-2-page-1.htm

Gathmann, C. & Schönberg, U. (2010). How General Is Human Capital? A Task-Based Approach. *Journal of Labor Economics*, 28(1), 1–49. https://doi.org/10.1086/649786

Gentzkow, M., Kelly, B. & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, 57(3), 535–574. https://doi.org/10.1257/jel.20181020

Guvenen, F., Kuruscu, B., Tanaka, S. & Wiczer, D. (2020). Multidimensional Skill Mismatch. *American Economic Journal: Macroeconomics*, 12(1), 210–244. https://doi.org/10.1257/mac.20160241

Hyman, B. G. (2018). Can Displaced Labor Be Retrained? Evidence from Quasi-Random Assignment to Trade Adjustment Assistance. In: *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association*, 111, 1–70.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, 76(4), 604–620. http://www.jstor.org/stable/1806062

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., ... & Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille. European Language Resources Association, 2479–2485. https://aclanthology.org/2020.lrec-1.302

Lindenlaub, I. & Postel-Vinay, F. (2021). The Worker-Job Surplus. NBER, *Working Paper* 28402. http://www.nber.org/papers/w28402

Marinescu, I. & Rathelot, R. (2018). Mismatch Unemployment and the Geography of Job Search. *American Economic Journal: Macroeconomics*, 10(3), 42–70. https://doi.org/10.1257/mac.20160312

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Şahin, A., Song, J., Topa, G. & Violante, G. L. (2014). Mismatch Unemployment. *American Economic Review*, 104(11), 3529–3564. https://doi.org/10.1257/aer.104.11.3529

Shaw, K. L. (1984). A formulation of the earnings function using the concept of occupational investment. *Journal of Human Resources*, 19(3), 319–340. https://doi.org/10.2307/145876

Traiberman, S. (2019). Occupations and import competition: Evidence from Denmark. *American Economic Review*, 109(12), 4260–4301. https://doi.org/10.1257/aer.20161925

Weston, J., Bengio, S. & Usunier, N. (2011). WSABIE: Scaling up to large vocabulary image annotation. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2764–2770. https://dl.acm.org/doi/10.5555/2283696.2283856

APPENDIX

ALTERNATIVE APPROACHES TO EVALUATING THE NEURAL NETWORK

Performance in Pretext Tasks

While training the neural network, we kept an eye on the variation in the general loss function defined above, as well as the capacity of the neural network to correctly predict the ROME code for a specific job offer (i.e. its accuracy).

These two values were calculated throughout the whole training and testing process, with training stopped when we found the parameters which delivered the highest level of accuracy with the training data set – around 80% in our case. We also found that unsupervised retraining of the language model on our corpus of job offers, before training our own neural network, significantly increased (by around 20%) the levels of accuracy we were able to attain. This suggests that the availability of large corpuses of job offers, such as the JOCAS database, can be invaluable when it comes to training models, even if the ROME codes corresponding to these jobs are not known (they can now be imputed by statistical learning techniques).

We observed a general reduction in the level of penalties during the training process. Breaking that down loss by loss, we found that, in spite of the normalisation, the task of predicting a ROME code on the basis of a job offer continued to play the most important role in training. The other losses dropped off fairly quickly, most likely because the constraints involved were easier to satisfy within the geometry of the space.

Qualitative Validation of the Measure of Distance Between Occupations

For the purposes of our qualitative analysis, we focused on the occupation most frequently associated with the jobseekers from the 14 sectors of activity defined in ROME. For each of these 14 reference occupations, we identified the 5 closest occupations according to our criteria, including some which are not listed as suggested mobility options in the ROME classification (see Online Appendices S3 and S4). This second list illustrates the capacity of our neural network to predict plausible professional transitions other than those it was trained on. The results suggest that not only does our measure succeed in identifying the mobility suggestions found in the ROME classification V3, but it also often manages to propose other suggestions which appear to be qualitatively coherent. This corroborates our initial suspicion that the suggestions found in the Mobilities section of the ROME classification are relatively limited in scope, and do not reflect the true range of pertinent professional mobility opportunities. Nevertheless, it is worth noting that our measure of distance between occupations is less consistent for those occupations which are under-represented in our corpus of offers, such as Music and singing (L1202). In the Online Appendix S2, we also introduce a visualization technique to explore the representation that preserves the pairwise cosine distance in the neighborhood of a given occupation while projecting the neighbouring occupations in a 2D space.