Beyond the Salary: Unpacking Job Attributes in French Ads with NLP

github.com/alfonsoawd/Project-NLPJobAds

Alfonso AWADALLA & Sofia VACA
ENSAE – École Polytechnique
May 2025

Abstract

Using over six million French job ads from the 2020 JOCAS dataset, this paper examines how employers advertise wages and non-wage job attributes. We adapt an English-language taxonomy using topic modeling and embeddings, then apply rule-based matching to identify the presence of specific amenities in job descriptions. We find that job ads are structured in systematically different ways depending on whether or not the wage is disclosed. And that industry fixed effects explain most of the variation in the attributes advertised. Our results suggest that employers use job ads strategically, and that non-monetary factors play a key role in signaling job quality and attracting applicants.

1 Introduction

The following paper studies the characteristics of job offers advertised by employers in France, with a particular focus on both wage and non-wage amenities. Using a dataset of online job postings that includes information such as location, posting dates, durations, industry and wage, we use the textual content of job descriptions to extract and identify the type of amenities offered.

The study has three main objectives. First, we aim to identify the most common wage and non-wage attributes that are mentioned by employers in job ads. Second, examine whether the presence of wage information is associated with differences in the attributes advertised. Third, we analyze if the variation in advertised attributes can be explained by other characteristics of the job add.

The paper is structured as follows. Section 2 provides contextual background and a review of the relevant literature. Section 3 presents the data and Section 4 outlines the empirical strategy, including the methodology used for identifying and classifying amenities. Section 5 discusses the main results, and Section 6 concludes with a summary of findings and a discussion of limitations.

2 Context

A growing strand of the literature explores the value of job attributes beyond wages, showing that workers may value non-wage amenities, such as flexibility, health coverage, or work environment—sometimes even more than a wage increase. Maestas et al. [2023] use stated-preference experiments to estimate workers' willingness to pay for these characteristics, highlighting how non-monetary factors can play a crucial role in job choices and overall job satisfaction.

Recent studies also focus on how employers communicate these attributes. Audoly et al. [2024] apply NLP techniques to classify pay and non-pay attributes in Norwegian job ads, finding that while 55% mention wages, nearly all refer to non-wage features. Firm-level fixed effects explain a large share of

Preprint. Under review.

variation in ad content, suggesting consistent posting styles. This is directly relevant to our work, which applies a similar approach to the French context.

Given the lack of a unified international taxonomy, ILO [2025] proposes a rule-based NLP method to extract non-wage amenities from online ads in emerging economies, covering dimensions such as flexibility, career development, and work-life balance. Along with Audoly et al. [2024], this provides a valuable framework for our classification strategy.

Other studies have applied NLP to detect specific amenities. Hansen et al. [2023] focus on flexible work arrangements and identify remote work offers, showing a threefold increase since 2019. Adams-Prassl et al. [2020] focus on employer-provided training, finding higher rates of this amenity in more concentrated labor markets. These studies show how NLP enables scalable analysis of job quality features that are otherwise hard to observe. Similarly, Deming & Kahn [2018] use NLP techniques in job vacancy texts to uncover firm-level variation in skill demands (both cognitive and social). Their findings highlight that job postings can reveal meaningful firm-level variation in skill demands.

Finally, search frictions literature underscores the role of job ads in signaling. Marinescu & Wolthoff [2020] find that job titles better predict applicant behavior than wages, while Sockin & Sojourner [2023] show that jobseekers actively seek firm-level insights. These studies support our focus on job ads as a central channel of information in the matching process.

3 Data

3.1 JOCAS Database

Our main data source is the JOCAS database (Job Offers Collection and Analysis System), developed by DARES to centralize daily job postings throughout France [DARES, 2023] [1]. We use the full 2020 dataset, totaling around 40 GB and comprising several thousand CSV files. Data-related limitations are discussed in the final section.

3.2 ROME and FAP Linking Table

The JOCAS database includes ROME codes, a highly detailed job classification system [2]. For simpler analysis and a broader comparison, we use the official correspondence table from DARES [2024] to map each ROME code to one of the 22 aggregated job categories, reducing the complexity of 531 distinct ROME codes.

3.3 Attribute Base

A core component of this study is the extraction of job attributes from job descriptions. As a baseline, we use the attribute base developed by Audoly et al. [2024], originally in English and tailored to the Norwegian context.[3] Notebooks 01, 02, and 03 produce the final version of the attribute base, *attribute_base_processedV3.xlsx*, available in the project's data folder.

4 Empirical Strategy

4.1 Extraction and Classification of Job Attributes

We implemented a four-step pipeline to extract and classify wage and non-wage attributes from French job ads. First, we translated the original English taxonomy from Audoly et al. [2024] into French using ChatGPT o4 (prompt detailed in `01-Attribute Base Translation.ipynb`), removed terms related to the Norwegian context (country of study in Audoly et al. [2024]), and enriched with morphological variants and french market-specific synonyms.

Second, to uncover missing terminology, we applied Latent Dirichlet Allocation topic modelling to a sample of 100 000 job ads (`02-Topic Modelling.ipynb`). The manual analysis of extracted

[1] More information on data collection methodology can be found here.

[2] More information on ROME here.

[3] The full attribute base is available here.

expressions per topic yielded two new categories—*ticket restaurants* and *chèques vacances*—and added approximately 50 expressions into existing attributes.

Third, to enrich the existing attribute expressions with similar but previously uncovered ones, we used word embeddings (see `03-Word Embedding.ipynb`) and followed a three-step process:

- **Seed extraction:** generated seed words per category by loading the updated attribute table, cleaning each expression (lowercasing, removing French stopwords and punctuation), and collecting the remaining tokens as the category's core vocabulary.
- **Candidate mining:** sampled a representative set of job ads, extracted all 3–7-word noun phrases via spaCy, and tagged any phrase containing a core vocabulary word as a candidate for that attribute category.
- **Embedding & selection:** embedded category's expressions with Sentence-BERT and averaged them to form a centroid vector per category; embedded all candidate expressions, ranked them by cosine similarity to the centroid, selected the top 30, and manually reviewed them to select relevant expressions.

This approach exploits SBERT's semantically meaningful embeddings — unlike the word2vec model used in Audoly et al. [2024] — to achieve better results and uncover hidden expressions [Reimers & Gurevych, 2019]. The analysis added approximately 100 expressions into existing attributes. The final attribute base was deduplicated using a clean, lemmatized form of the expressions to match the preprocessing applied to job-ad texts.

Finally, we applied rule-based pattern matching to the full JOCAS 2020 corpus (`04-Jocas Information Extraction.ipynb`). Matches are disregarded if a keyword is preceded or followed by negations such as "pas de", "non". This produced a structured dataset where each row represents a job ad, with variables such as location, FAP category, and binary columns for each attribute and aggregated category (1 if the attribute is mentioned, 0 otherwise).

4.2 Econometric Analysis of Amenities Mentions

We first estimate a series of logistic regressions to examine whether the presence of wage information is associated with differences in the attributes advertised. Each regression takes a job attribute as a binary outcome, with the main explanatory variable being whether the ad explicitly mentions the wage. These models also include fixed effects for industry and location as control variables, given their potential correlation with both wage disclosure and the presence of specific amenities.

In a second series of regressions, inspired by Audoly et al. [2024], we adopt a stepwise modeling strategy to better understand the relative contribution of different factors in explaining the presence of advertised attributes. In this case, we incrementally introduce fixed effects for industry, location, and wage disclosure. This specification allows us to assess the added explanatory power of each variable, using McFadden's pseudo-R2 as a measure of model fit.

For both sets of regressions, and given the large size of our dataset, we collapsed the data by unique combinations of the outcome and explanatory variables. Each combination was assigned a frequency weight corresponding to the number of repeated observations. The models were estimated using a Generalized Linear Model (GLM) with a logit link and binomial family, allowing us to incorporate the frequency weights and work with the full sample without the need for random sub-sampling.

5 Results

5.1 Descriptive Patterns of Job Attributes

Figure 1 shows the prevalence of job attributes across postings on different online job platforms in France in 2020. The most frequently advertised group of attributes is Task-related attributes which is present in 70% of the job ads. Within this category, the most common attributes mentioned are: Responsibilities in the job and autonomy in the execution of tasks.

Contract duration and minor advantages follow, each appearing in over 30% of ads. Within Contract duration, the permanent job attribute is the most mentioned. Between 20-30% of the ads mention career opportunities and around 20% mention the regime scheme. Advantageous schedules are

mentioned in around 18% of the ads. In contrast, financial attributes and demanding schedules are much less frequently mentioned.

Overall, the high prevalence of task- and contract-related attributes suggests that employers focus on describing the nature of the role and its stability. On the other hand, the low frequency of salary-related mentions may reflect a deliberate omission. In the next section, we examine whether the presence of wage information is associated with differences in the attributes advertised.

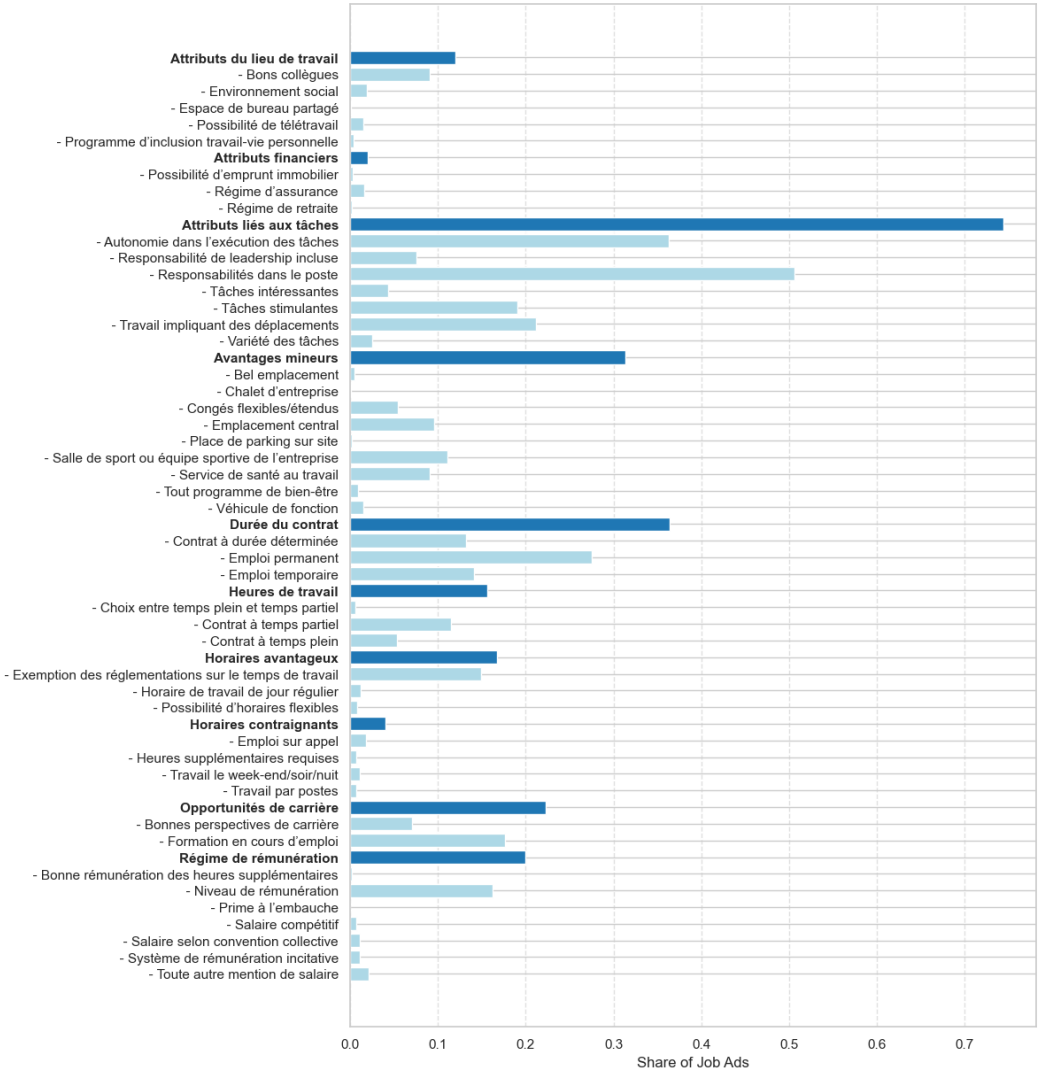

Figure 1: Prevalence of Job Attributes in Job Ads

5.2 Job attributes and wages

To assess whether job postings that mention wages differ in the attributes they highlight compared to those that don't, we first compared the share of amenities advertised in each group. Figure 2 shows that postings with wage information are more likely to include task-related attributes, while those without wage disclosure tend to emphasize contract duration, type of remuneration, minor advantages, and career opportunities.

To ensure the descriptive results aren't simply reflecting differences across industries or locations, which could influence both the likelihood of mentioning wages and the attributes included, we estimated a logit model with fixed effects for industry and location. Figure 3 shows that the patterns remain: postings that mention wages are more likely to emphasize job tasks, while those that don't

tend to highlight contract stability and financial aspects. This suggests that employers who choose not to disclose wages (potentially because they're lower) may use other benefits to attract applicants, such as contract duration, remuneration type, and minor advantages. Overall, there can be seen a significant difference among the attributes that are advertised between the ads that specify wages and the ads that don't.

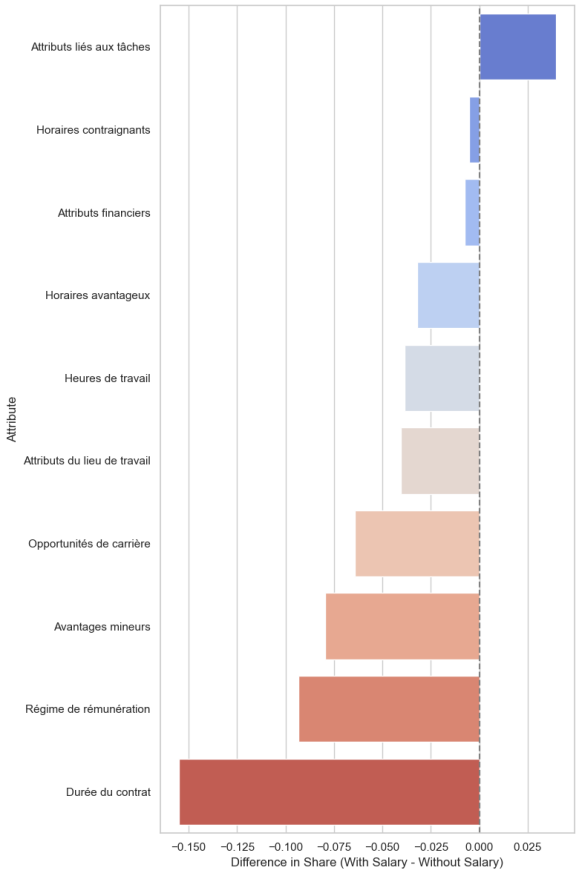

Figure 2: Significant Differences in Attribute Prevalence (With vs. Without Salary Disclosure)

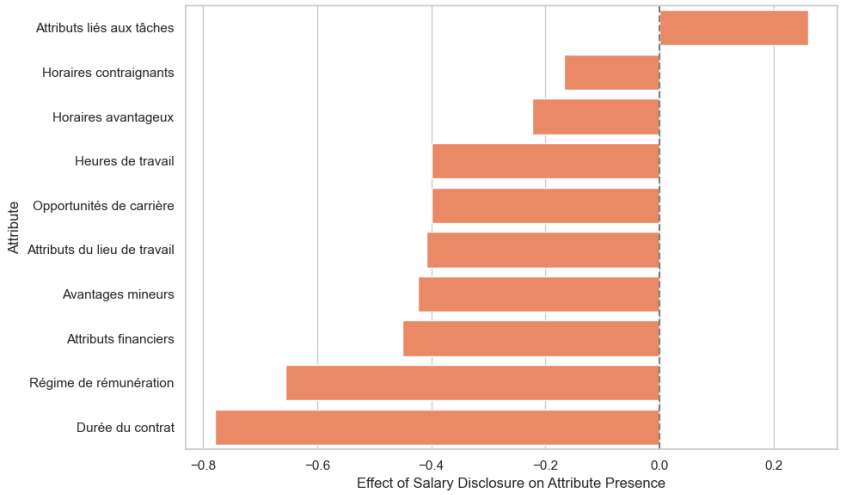

Figure 3: Logit Coefficient for Salary Disclosure

5.3 Drivers of Non-Wage Amenities

Figure 4 and Figure 5 present the results from the second series of logistic regressions, where fixed effects are incrementally introduced to explain the presence of job attributes. The stacked bars reflect the relative contribution of each dimension.

Figure 4 shows that most of the variation in the presence of job attributes is explained by industry differences, especially for financial and working time-related attributes. Location contributes marginally, while wage specification adds explanatory power for attributes like contract duration, remuneration schemes, and working hours. Figure 5, which presents results at the sub-attribute level, confirms this pattern. Industry fixed effects explain most of the variation in benefits and working conditions. And wage disclosure, contributes less overall, but still explains meaningful variation in a number of specific amenities.

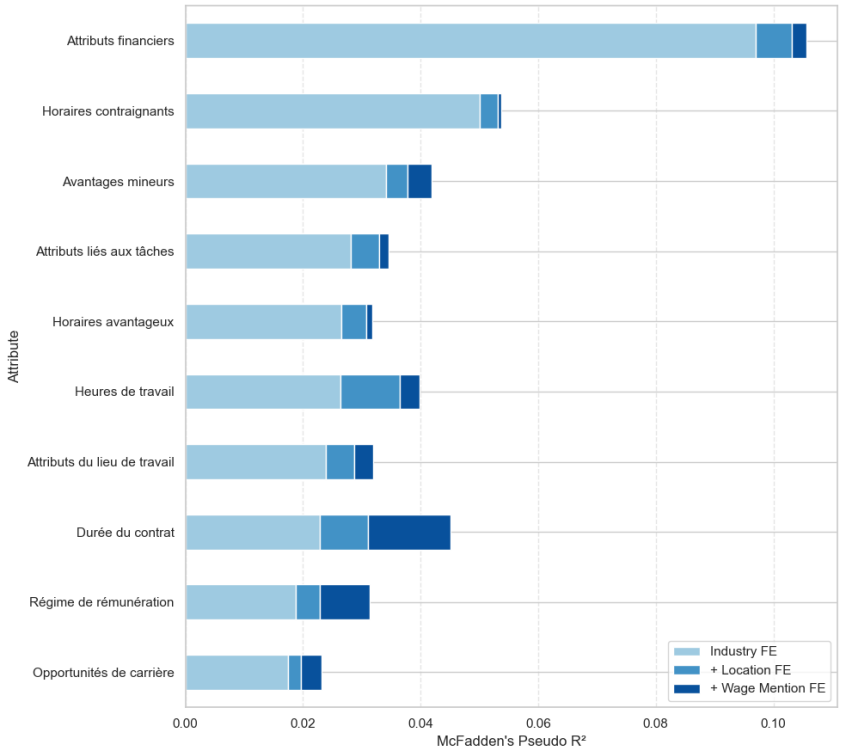

Figure 4: Explained Variation in Aggregated Attributes by Fixed Effects

6 Conclusion

This paper analyzed job postings in France to understand how employers advertise wage and non-wage amenities. First, we identified that task-related attributes are the most frequently mentioned, followed by contract duration and minor advantages. Second, we found that job ads are structured in systematically different ways depending on whether or not the wage is disclosed. Third, using logistic regressions with fixed effects, we showed that industry explains much of the variation in advertised attributes, but wage disclosure also contributes independently—especially for contract and compensation-related elements. Overall, our results suggest that employers use job ads strategically, with or without wage information, to signal job quality. This shows that, non-monetary factors play a key role in shaping perceptions and job choices, consistent with findings in the literature.

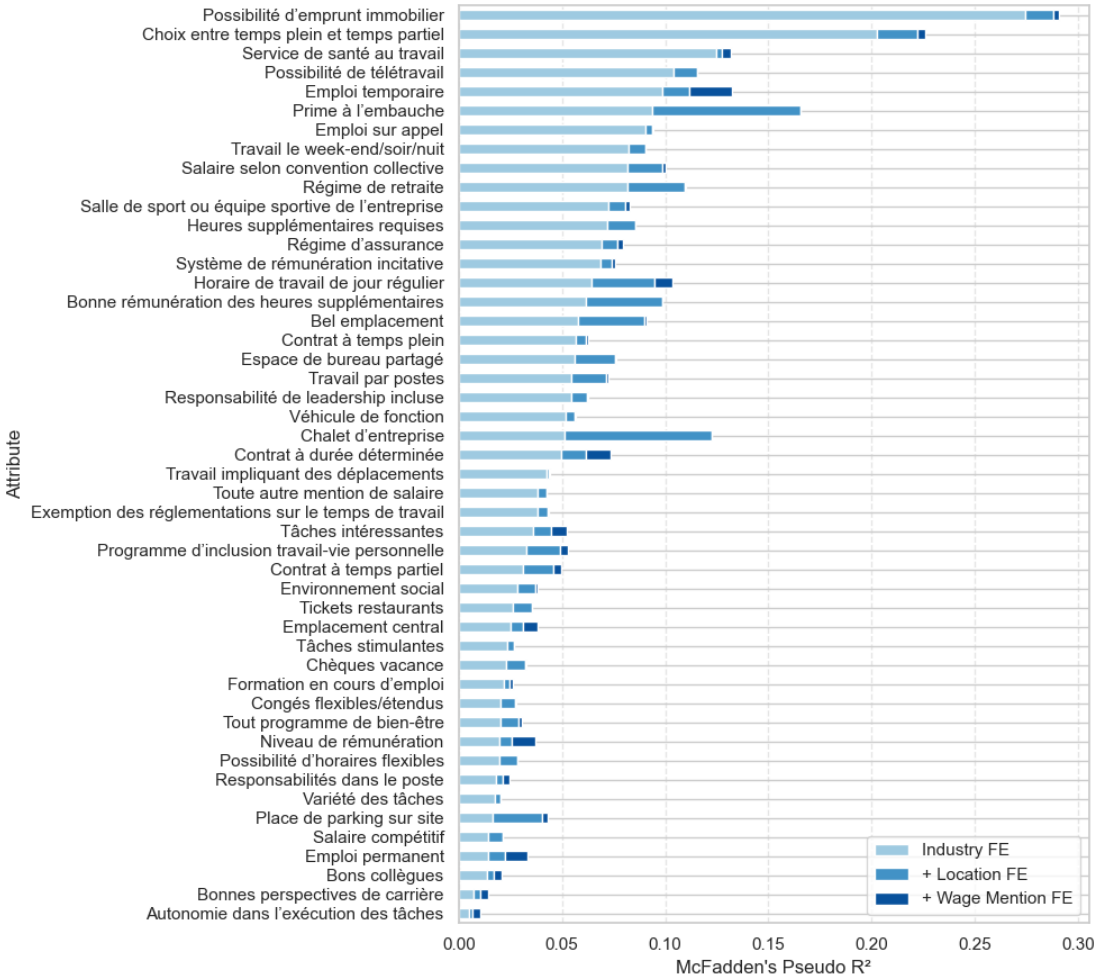

Figure 5: Explained Variation in Attributes by Fixed Effects

6.1 Limitations

- ***Bias inherent to JOCAS data:*** Job descriptions may omit key details like contract type or salary, which often appear in separate sections/boxes, thus not captured by our analysis. However, the `is_salary` variable is based on the full job ad and should be reliable. JOCAS also overrepresents managerial and online-recruited roles, while sectors with mass or informal hiring are underrepresented.
- ***Bias due to incomplete extraction:*** Due to time and computing constraints, the attribute extraction loop (notebook `04`) ran for two days and processed just over half of the job ads (over 6 million rows). While the exact bias is unknown, the large sample size provides reasonable confidence in the robustness of the results.
- ***Semantic bias and hidden information:*** Our method relies on surface-level textual patterns and may overlook implicit or context-dependent expressions. Although we mitigated this limitation by enriching our attribute base, some attributes may still be expressed in ways that fall outside our extracted vocabulary.

References

Audoly, R., Bhuller, M., & Reiremo, T. A. (2024). The Pay and Non-Pay Content of Job Ads. *arXiv preprint* arXiv:2407.13204v2. . NeurIPS.

Snell, J., Swersky, K., & Zemel, R. S. (2017). Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, 30.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

DARES. (2023). *Base JOCAS – Document méthodologique résumé*. `https://dares.travail-emploi.gouv.fr/publication/`. Accessed April 2025.

DARES. (2024). *La nomenclature des familles professionnelles 2021*. https://dares.travail-emploi.gouv.fr/donnees/la-nomenclature-des-familles-professionnelles-2021. Accessed March 2025.

Maestas, N., Mullen, K. J., Powell, D., von Wachter, T., & Wenger, J. B. (2023). *The Value of Working Conditions in the United States and the Implications for the Structure of Wages*. American Economic Review, 113(7), 2007–2047. `https://doi.org/10.1257/aer.20210415`. Accessed April 2025.

ILO. (2025). *Measuring Quality of Employment in Emerging Economies: A Methodology for Assessing Job Amenities Using Online Data*. ILO Methodological Brief No. 2. `https://doi.org/10.2139/ssrn.3957002`. Accessed April 2025.

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R., & Taska, B. (2023). *Remote Work across Jobs, Companies, and Space*. NBER Working Paper No. 31007. `https://www.nber.org/papers/w31007`. Accessed April 2025.

Adams-Prassl, A., Balgova, M., & Qian, M. (2020). *Flexible Work Arrangements in Low Wage Jobs: Evidence from Job Vacancy Data*. IZA Discussion Paper No. 13691. `https://www.iza.org/en/publications/dp/13691`. Accessed April 2025.

Deming, D., & Kahn, L. (2018). *Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals*. Journal of Labor Economics, 36(S1), S337–S397. `https://doi.org/10.1086/694106`. Accessed April 2025.

Marinescu, I., & Wolthoff, R. (2020). *Opening the Black Box of the Matching Function: The Power of Words*. Journal of Labor Economics, 38(2), 535–568. `https://doi.org/10.1086/705880`. Accessed April 2025.

Sockin, J., & Sojourner, A. (2023). *What's the Inside Scoop? Challenges in the Supply and Demand for Information on Employers*. Journal of Labor Economics, 41(4), 1041–1079. `https://doi.org/10.1086/722936`. Accessed April 2025.