



HAL
open science

Designing and Evaluating Labor Market Recommender Systems

Elia Pérennès

► **To cite this version:**

Elia Pérennès. Designing and Evaluating Labor Market Recommender Systems. Economics and Finance. Institut Polytechnique de Paris, 2024. English. NNT : 2024IPPAG011 . tel-04909305

HAL Id: tel-04909305

<https://theses.hal.science/tel-04909305v1>

Submitted on 23 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2024IPPAG011

Thèse de doctorat



Designing and Evaluating Labor Market Recommender Systems

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°626 : École Doctorale de l'Institut Polytechnique de Paris (ED IP
Paris)

Spécialité de doctorat : Sciences économiques

Thèse présentée et soutenue à Palaiseau, le 10 décembre 2024, par

ELIA PÉRENNÈS

Composition du Jury :

François Fontaine Professeur des universités, Université Paris 1	Président
Michèle Belot Professeure, Cornell University	Rapporteure
Rafael Lalive Professeur, Université de Lausanne	Rapporteur
Roland Rathelot Professeur Associé, ENSAE (CREST)	Examineur
Alexandra Roulet Professeure Assistante, INSEAD	Examinatrice
Bruno Crépon Professeur, ENSAE (CREST)	Directeur de thèse
Cyril Nouveau Directeur DSEE, France Travail	Invité

Remerciements

Ce manuscrit de thèse n'a pu voir le jour que grâce au soutien des nombreuses personnes qui m'ont accompagnée tout au long de ce parcours. Je tiens d'abord à exprimer ma gratitude à Bruno Crépon, mon directeur de thèse, pour son accompagnement tout au long de ce cheminement. Avant de le rencontrer lors de ma deuxième année de master, je n'avais jamais envisagé poursuivre en thèse. Au contraire, l'idée même d'un doctorat, à moins qu'il ne soit orienté vers des applications concrètes, ne m'attirait pas. Il a su éveiller en moi l'envie de me lancer dans ce défi en me proposant un projet en contrat CIFRE avec France Travail, ce qui correspondait parfaitement à mes aspirations. Tout au long de cette aventure, Bruno a fait preuve d'une disponibilité et d'une accessibilité remarquables, partageant généreusement sa grande expertise en économétrie lors de nos nombreux échanges. Son enthousiasme communicatif et sa passion pour l'évaluation des politiques publiques sont pour moi une grande source de motivation.

Je suis très honorée que Michèle Belot, François Fontaine, Rafael Lalive, Cyril Nouveau, Alexandra Roulet et Roland Rathelot aient accepté de faire partie de mon jury de thèse et je les remercie sincèrement. Merci en particulier à Michèle Belot et à Rafael Lalive qui ont tous deux accepté d'être les rapporteurs de cette thèse.

Je voudrais également remercier chaleureusement mes co-auteurs des équipes Vadore et ASP : Guillaume Bied, Philippe Caillou, Bruno Crépon, Christophe Gaillac, Morgane Hoffmann, Solal Nathan, Mitia Oberti, Michèle Sebag, Abhijit Banerjee et Cécile Welter-Médée. Avec l'équipe Vadore, j'ai eu l'opportunité de mener un projet ambitieux et exigeant, et je suis reconnaissante d'avoir partagé cette aventure avec une équipe aussi engagée et compétente. Un immense merci en particulier à Guillaume Bied, mon binôme sur ce projet, dont les avis toujours pertinents, la bienveillance et la rigueur sont d'une valeur inestimable. J'espère que nous aurons l'occasion de relever encore de nombreux défis ensemble. Je remercie également Morgane pour tous les moments, aussi bien les bons que les plus difficiles, que nous avons partagés au CREST, à France Travail et ailleurs. Sa gentillesse, son énergie positive et sa détermination ont été une source de soutien précieuse. Enfin, un grand merci à Cécile, avec qui il a été si agréable de travailler. Je suis heureuse d'avoir croisé le chemin de ces personnes au cours de ce parcours.

Je suis profondément reconnaissante envers les équipes de la Direction Générale de France Travail pour leur accueil chaleureux et la confiance qu'ils m'ont témoignée, ce qui a permis à ma thèse de se dérouler dans des conditions idéales. Je remercie en particulier Cyril Nouveau, qui m'a intégrée au sein de la DSEE avec une ouverture rare à la recherche et à l'expérimentation, facilitant ainsi l'avancement de mes travaux en leur accordant une pleine légitimité. Je suis également reconnaissante envers les collègues de France Travail qui ont largement contribué à la mise en œuvre des expérimentations réalisées dans le cadre de ma thèse, notamment Hélène Caillol, Peggy Duhayon, Sébastien Robidou et Chantal Vessereau. Un grand merci également à mes camarades de la team Big Data avec qui j'ai le plaisir de collaborer chaque jour sur des projets très stimulants : Paul, Pierre-Antoine, Yann et Julia. Toujours unis, nous affrontons les défis avec joie et bonne humeur ! Merci aussi à mes (ex-)co-bureaux qui font passer les jours au bureau en un éclair : Christophe, Jérôme, Laura, Mitia et Yohan. Merci enfin aux autres collègues de la DSEE avec qui j'ai le plaisir de partager des bons moments au quotidien et qui contribuent à créer un environnement agréable et stimulant (la liste est longue et j'en oublie sûrement) : Alexandre, Amine, Andréa, Anne, Aurélien, Chantal, Clara, Danielle, Eric, Fanny, Greg, Hélène, Jacques, Joseph, Lucile, Mamadou, Manon, Marc, Marie, Michel, Mylène, Nicolas V, Oumar, Snjezana, Sophie, Vicente, Yannick et Yves.

J'ai également trouvé un environnement de travail très agréable au sein du CREST, pour lequel je remercie chaleureusement tout le personnel administratif. Ensuite, je suis particulièrement reconnaissante envers Roland Rathelot, dont l'expérience et la sagesse m'ont souvent été d'un grand secours. Mes remerciements s'étendent à Benoit Schmutz et Arne Uhlendorff, les autres membres de mon comité de suivi, dont les conseils avisés ont été très utiles à ma progression. Je souhaite bien sûr remercier également les doctorants du CREST pour les moments de convivialité que nous avons partagés. Je tiens à remercier particulièrement les ex-membres du bureau 4094 : Guidogiorgio, Étienne, Pauline, et mes deux acolytes, Esther et Émilie. Je voudrais aussi adresser un clin d'œil chaleureux à mes camarades CRESTINS : Rémi, mon gars sûr depuis le collège MLK, Gwen-Jiro, le Breton le plus chauvin que je connaisse, Jules, le plus beau bébé du monde, Étienne, le Cadurcien toujours serein, et Fabien, un ami au quotient émotionnel exceptionnel. Je tiens également à remercier tous les autres doctorants et stagiaires que j'ai eu le plaisir de croiser lors de mes presque quatre années passées au CREST, parallèlement à mon temps chez France Travail. Merci à Alexis, Anasuya, Antoine B, Antoine F, Bérengère, Christophe, Denys, Elio, Germain, Ivan, Jérôme, Jérémy H, Jérémy L, Juliette, Léa B, Louis-Daniel, Lucas, Martin, Morgane C, Morgane G, Pauline C, Pierre-Edouard, Reda, Tang et Thomas D.

Ce manuscrit marque la fin d'un parcours exigeant. Je suis profondément reconnaissante envers celles et ceux qui ont éclairé ce chemin de leur soutien et de leur présence.

Merci infiniment à mes amis Alexis, Anne, Basile, David, Hervé, Julien, Laura, Loïc, Marc, Marie, Nagui, Nilai, Opale, Pauline, Solène, Ségolène, Thomas, et aussi à mes deux filleules adorées Ambroise et Chloé. Je tiens à remercier également ma famille, en particulier mes parents, Ronan et Anne-Marie, mes frères Mathis et Milo, et mes grands-parents Alain, Annick et Yvette, qui m'ont toujours soutenue et ont cru en moi tout au long des hauts et des bas de cette thèse.

Mon plus grand merci est pour Pierre. Il m'est impossible de trouver les mots pour lui exprimer toute ma gratitude, tant son soutien a été essentiel ces dernières années. Enfin, un immense merci à notre fille Gabrielle, dont la malice et la joie communicative ont illuminé les derniers moments de ce parcours. Je lui dédie cette thèse, avec tout mon amour.

Résumé

L'essor des systèmes de recommandation d'offres d'emploi représente un enjeu majeur, avec la promesse de transformer les marchés du travail en optimisant le processus d'appariement entre les demandeurs d'emploi et les offres. Pour exploiter pleinement ce potentiel, des questions doivent néanmoins être adressées concernant leur conception, leur personnalisation, et leur acceptation par les utilisateurs. Cette thèse se penche sur ces enjeux à travers trois chapitres explorant chacun un de ces aspects des systèmes de recommandation d'offres d'emploi.

Le premier chapitre explore comment la théorie économique peut aider à identifier un système de recommandation (SR) optimal du point de vue des demandeurs d'emploi (DE). Nous introduisons le chapitre en décrivant et comparant deux SR typiques. Le premier, que nous appelons \mathcal{U} (pour *utilité*), est couramment utilisé par les Services Publics de l'Emploi (SPE). Il évalue la compatibilité entre les offres d'emploi et les DE en mesurant la proximité entre le profil et les critères de recherche du DE, et les caractéristiques des offres. Le second SR, que nous appelons \mathcal{P} (pour *probabilité*), est inspiré d'une littérature récente sur l'apprentissage automatique. Il prédit les appariements entre DE et offres ayant le plus de chances de mener à une embauche, en analysant les caractéristiques des deux parties. Nous montrons que ces deux SR génèrent effectivement des recommandations très différentes. Nous proposons ensuite un modèle théorique économique pour formaliser la décision de postuler aux offres des DE, en se concentrant sur deux variables clés : l'utilité de l'emploi (U) et la probabilité de succès de la candidature (P). En utilisant des données sur les clics et les candidatures effectuées par les DE, nous montrons que U et P influencent significativement les décisions de candidature des DE, conformément aux enseignements du modèle. Ce modèle révèle également que le SR optimal du point de vue des DE intègre les deux dimensions U et P dans un score unique. Enfin, sur le même jeu de données, nous comparons empiriquement les ensembles de recommandations issus de différents SR — dont ceux basés sur les scores d'utilité (\mathcal{U}) et de probabilité de succès d'une candidature (\mathcal{P}), ainsi que le SR optimal — afin d'évaluer leur valeur ajoutée par rapport au comportement de recherche organique des DE (i.e. ce qu'ils font naturelle-

ment en l'absence de recommandations). Nos résultats montrent que le SR optimal et le SR basé sur la probabilité d'embauche (P) apportent une valeur ajoutée notable par rapport aux recherches organiques effectuées par les DE, alors que le SR basé sur l'utilité (U) est moins efficace. De plus, nous observons que l'impact du SR optimal varie selon les individus, favorisant davantage ceux proches du marché du travail. Au total, notre étude met en évidence l'importance d'identifier ces deux quantités, essentielles pour concevoir un SR : l'utilité d'un emploi (U) et la probabilité de succès d'une candidature (P). Cette dernière peut être estimée à partir des données disponibles, domaine dans lequel les outils d'apprentissage automatique excellent. En revanche, prédire l'utilité d'un poste pour les DE, bien que cela paraisse plus simple, reste complexe en raison de l'absence d'observations mesurant directement cette grandeur.

Un des défis majeurs dans la conception des SR pour le marché du travail, comme souligné dans le premier chapitre, réside dans la complexité de mesurer l'utilité que les DE retirent des offres d'emploi. Cette question se prolonge et se précise dans le second chapitre où nous explorons l'impact de l'intégration des préférences des DE concernant divers attributs des emplois au sein du SR du SPE français. Ce SR (similaire au SR U étudié dans le chapitre 1) analyse une multitude de données sur les DE et les offres d'emploi, incluant qualifications, expérience, attentes salariales, localisations, horaires, etc. Le SR calcule des correspondances entre ces critères (par exemple, entre le salaire proposé et les attentes salariales ou entre le métier recherché et offert) et leur attribue des poids spécifiques pour mesurer la compatibilité globale entre DE et offre. Les offres avec les scores d'adéquation les plus élevés sont proposées aux DE. Néanmoins, ce système souffre d'une rigidité notable en matière de personnalisation : il applique de manière uniforme les mêmes pondérations pour tous, sans prendre en compte les variations individuelles des priorités professionnelles. Dans ce contexte, nous conduisons un essai contrôlé randomisé où les DE participants répartissent 15 points entre cinq attributs des offres d'emploi – la profession, le salaire, la proximité domicile-travail, le type de contrat et le nombre d'heures par semaine – pour définir leurs préférences personnelles. Les données ainsi récoltées mettent en lumière des divergences notables avec les pondérations standards, les participants valorisant davantage des aspects comme le salaire et la proximité domicile-travail plutôt que la profession. Ensuite, les résultats de notre expérimentation montrent que la personnalisation des recommandations en fonction des préférences individuelles entraîne une hausse significative de 13,8% des clics sur les recommandations et de 23,1% des candidatures comparativement à l'algorithme standard non personnalisé, grâce principalement à la mise en avant de nouvelles offres que les DE n'auraient pas vues autrement. Toutefois, comparé à l'algorithme de base sans recueil de préférences, l'effort

supplémentaire demandé pour recueillir ces préférences cause une réduction de 26% des clics sur les recommandations et de 30% des candidatures. Ces observations suggèrent que, si la personnalisation peut effectivement améliorer la pertinence des recommandations, l'effort supplémentaire requis peut décourager l'engagement des utilisateurs, soulignant ainsi le besoin de trouver un juste milieu entre les bénéfices de la personnalisation et les contraintes qu'elle impose.

Enfin, le chapitre 3 explore une dimension davantage comportementale en analysant la manière dont les utilisateurs perçoivent et interagissent avec les systèmes de recommandation d'offres d'emploi. Nous mettons en œuvre un essai contrôlé randomisé à grande échelle pour examiner l'impact de la présentation des recommandations sur l'engagement des DE. Dans cette expérimentation, les DE reçoivent des recommandations personnalisées issues d'un unique SR et sont répartis en trois groupes de traitement. Dans chaque groupe, les recommandations sont présentées différemment : soit clairement identifiées comme issues d'un algorithme, soit comme élaborées grâce à l'expérience d'humains, soit sans aucune indication sur leur origine. Cette méthode permet d'évaluer les réactions des DE en fonction uniquement de la manière dont les recommandations sont présentées. Les participants évaluent initialement cinq offres d'emploi avant de décider s'ils souhaitent les explorer plus en détail et éventuellement postuler. Les résultats montrent une aversion notable pour les recommandations marquées explicitement comme algorithmiques : les DE montrent un intérêt et un taux de clics significativement inférieurs par rapport aux recommandations présentées comme étant d'origine humaine. Cette réticence persiste même si les recommandations sont étroitement alignées avec leurs préférences, indiquant que la qualité perçue des recommandations ne suffit pas à elle seule pour contrer les attitudes négatives engendrées par les sources algorithmiques. Pour approfondir notre compréhension de cette aversion, nous examinons comment certaines caractéristiques des utilisateurs influencent leurs réactions aux recommandations. En utilisant des techniques avancées d'apprentissage automatique, nous identifions des sous-groupes d'individus avec des attitudes diamétralement opposées envers les technologies de recommandation. Le sous-groupe que nous qualifions de "favorable aux algorithmes" est principalement constitué de jeunes DE et de personnes en chômage de longue durée, qui démontrent un engagement plus marqué face aux recommandations algorithmiques. De plus, ces individus bénéficient souvent d'un accompagnement plus intensif par le SPE et sont davantage actifs sur des marchés de l'emploi tendus (avec plus d'offres d'emploi que de DE). À l'opposé, le groupe "averse aux algorithmes", composé majoritairement de DE plus âgés, avec des attentes salariales élevées et des périodes de chômage plus courtes, montre une préférence pour les recommandations présentées comme humaines. Ce

groupe évolue généralement sur des marchés moins favorables et bénéficie d'un soutien moins intensif de la part du SPE. Ces différences marquées soulignent la complexité des réactions face aux algorithmes de recommandation et l'importance de personnaliser la présentation des SR pour pouvoir maximiser leur impact.

Contents

Introduction	17
1 Designing Job Recommender Systems:	
How to Improve Human-Based Search	46
1 Study context	52
1.1 Review of existing job RSs	52
1.2 A machine learning RS based on hiring predictions	54
1.3 A preference-based RS based on search criteria	56
1.4 Data	57
2 A first look at the recommendations generated by the two RSs	59
2.1 Evaluating RSs predictive performance	59
2.2 Estimation of the matching probability using the ML score as predictor	60
2.3 Contrasting preference-based and ML-based rankings	62
2.4 Assessing the potential of combining the two RSs in the field	65
2.5 Overall takeaway	69
3 Job search model with RSs	70
3.1 Model framework	71
3.2 Searching in absence of the RS	72
3.3 Identification of Γ	74
3.4 Search using a RS	75
3.5 Empirical validation of the model	77
4 Empirical evaluation of the different RSs	80
4.1 Evaluation metrics	81
4.2 Comparative performance of the RSs	83
Appendices	96
A Field experiment	97
B Model appendix	106
C Additional figures and tables	107

2 Personalization Pitfalls: The Unintended Effects of Using Stated Preference

Data in Job Recommendations	112
1 Context and study design	119
1.1 Institutional context and the PES's current job recommender system	119
1.2 Study design	122
1.3 The online platform	129
1.4 Technical challenges and unanticipated issues	132
2 Data and samples statistics	134
2.1 Data	134
2.2 Description of the samples of participants	135
3 Results from the pilot study: evaluating the reliability of the preference measurement	139
3.1 Respondents' engagement with the survey tasks	139
3.2 Are preferences revealed through the ranking consistent with those obtained through the point allocation question?	140
3.3 Which weights fit better revealed preferences: standard or personalized?	144
4 Analysis of stated preferences for job attributes	145
4.1 Average preferences for job attributes	145
4.2 Typical preferences profiles	146
5 Descriptive analysis of the recommendations generated by the two matching algorithms	149
5.1 How does weight personalization affect the recommendations made?	149
5.2 How did the implementation issue affect the intensity of our treatment?	152
6 Results of the large scale experiment	154
6.1 Effect of personalized recommendations compared to standard recommendations	154
6.2 Evaluating the impact of the stated preference data collection itself .	156
6.3 Overall effects of personalized job recommendations using stated preferences	158
Appendices	167
A The PES matching algorithm	167
B Metrics to compare lists of recommended ads	171
C User flows on the online interface	173
D Screenshots of the user interface, translated from French	175
E Descriptive statistics	181

F	Results from the pilot study	184
G	Analysis of stated preferences	185
H	Descriptive analysis of the recommendations	187
I	Results of the large scale experiment	189
3	Job Seekers' Responses to AI Job Recommendations: Insights from a Field Experiment	194
1	Field experiment	201
1.1	Intervention	201
1.2	Experimental design	203
2	Data and sample statistics	205
2.1	Data	205
2.2	Descriptive statistics	207
3	Results	210
3.1	On average, job seekers have algorithm aversion	210
3.2	Exploring heterogeneous effects by recommendation quality	213
3.3	Systematically identifying algorithm-averse and algorithm-friendly job seekers: a machine learning approach	215
	Appendices	239
A	The online survey	239
B	Summary statistics	243
C	Complements on results	248

List of Figures

1.1	High-level Structure of the neural network	56
1.2	Performance on the test set of different RSs.	60
1.3	Comparison of the rankings of the best recommendations wrt the other ranking	64
1.4	Comparison of the best recommendations in the two rankings: hiring probabilities and preference score	65
A.1	Scores distributions	99
A.2	Landing page	101
A.3	First page	102
A.4	Second page	103
2.1	Distributions of deviations from ideal DCG	142
2.2	Propensity to do better than random, according to potential gain	143
2.3	DCG of jobseeker’s ranking: personalized weights. vs. standard weights	145
2.4	Distributions of weights collected and ad-hoc weights used in the PES matching algorithm	146
2.5	Average weights by cluster, compared to average weights and default weights	147
2.6	Estimated effect of each covariate on each cluster’s belonging	149
B.1	Inverse of $\log_2(1 + x)$	171
C.2	User flow - pilot study	173
C.3	User flow - large scale experiment	174
D.4	Pilot study, invitation email	175
D.5	Large experiment, invitation email	176
D.6	Homepage	177
D.7	First screen	178
D.8	Second screen	179
D.9	Third screen	180
F.10	Distribution of time spent on each survey step	184
G.11	Distance between individual weights and equal weights	185

H.12	Maximum rank among clicks	187
H.13	Distributions of realized and potential overlaps	188
3.1	Conversion funnel	210
3.2	Group Average Treatment Effects	224
3.3	GATES on interest rate	224
3.4	GATES on click rate	224
3.5	Classification analysis on Click rate Recommendation quality	227
3.8	Classification analysis on Click rate User characteristics	228
3.11	Classification analysis on Click rate Unemployment spell characteristics	230
3.14	Classification analysis on Click rate Labor market conditions	232
A.1	Landing page	239
A.2	First page	240
A.3	Bottom of the first page	241
A.4	Second page	242
B.5	Association between declared interest and click behavior	247
C.6	Strength of the association between the CLAN covariates	251
C.7	Classification analysis on Interest rate Recommendation quality	252
C.8	Classification analysis on Interest rate User characteristics	253
C.9	Classification analysis on Interest rate Unemployment spell characteristics	254
C.10	Classification analysis on Interest rate Labor market conditions	255

List of Tables

1.1	Some examples of different RSs	54
1.2	Estimates of the best logistic predictor given the ML score	63
1.3	Treatment differences in recommendations appreciation	70
1.4	Estimates of the model of application on job postings	79
1.5	Average expected value of the different RSs	84
1.6	Comparison of the expected value of the different RSs	86
1.7	Similarity rates on recommended vacancies	87
A.1	Balance check among full sample	104
A.2	Survey completion	105
C.3	Information on offers and job seekers respectively used by the preference-based and machine-learning based RSs	108
C.4	Descriptive statistics on the populations with clicks	109
C.5	Estimates of the model of application on job postings	109
C.6	Average hiring rate of the different RSs	110
C.7	Comparison of the hiring rate of the different RSs	110
C.8	Shares of clicks and recommendations inside the declared occupational search area	111
2.1	Criteria and weights used in the standard PES matching algorithm	120
2.2	Treatment groups	129
2.3	Number of vacancies online as of April 26, 2022 (before randomization)	130
2.4	Assigned treatment vs. received treatment for participants who logged in	138
2.5	Treatment differences in participation and issues	139
2.6	Summary statistics on recommendations sets	151
2.7	Matching scores of recommendation sets	152
2.8	Impact of the personalization treatment on recommendations appreciation and adoption	155
2.9	Impact of being assigned to the partial treatment group	157
2.10	Impact of being assigned to the full treatment group	159

E.1	Comparison of full sample and participants to the pilot study	181
E.2	Summary statistics among full sample	182
E.3	Comparison of full sample and participants to the experiment	183
G.4	Average characteristics of each cluster vs. all respondents	186
I.5	First stage regressions: Probability of receiving full and partial treatments .	190
I.6	Impact of receiving the full treatment	191
I.7	Impact of receiving the partial treatment	192
I.8	Impact of the personalization treatment on recommendations appreciation and adoption	193
3.1	Treatment differences in participation	208
3.2	Impact of perceived source on the appreciation of recommendations, ex- perts as reference	212
3.3	Heterogeneous treatment effects by recommendation quality	215
3.4	Comparison of ML methods	222
3.5	Best Linear Predictor	223
3.6	Group Average Treatment Effects	224
3.7	Group Average Treatment Effects (quintiles for interest on clicks)	225
B.1	Balance check among full sample	243
B.2	Balance check among full sample	244
B.3	Comparison of full sample and participants to the experiment	245
B.4	Comparison of full sample and participants to the experiment	246
C.5	Impact of perceived source on the appreciation of recommendations, ex- perts as reference, negative binomial model	248
C.6	Flows between quintiles	250

Introduction

The rapid rise of job Recommender Systems (RS) presents urgent and complex challenges that demand attention. These systems hold the promise of reshaping labor markets by improving job matching, but realizing this potential requires answering important questions about their design, personalization, and user perception. This thesis explores these critical issues through three chapters, each evaluating a distinct dimension of RS design for the labor market.

This thesis is the result of a CIFRE contract¹ with France Travail, the Public Employment Service in France. This collaboration offers several advantages: it enables direct engagement with the operational challenges faced by employment services, it facilitates the design and execution of ambitious projects, such as the development and testing of RSs in real-world settings, and supports the implementation of large-scale randomized controlled trials (RCTs).

One of the key projects that shaped this work is Vadore^{2,3}. Vadore is a collaborative research project that integrates expertise from economics and computer science. In addition to its partnership with France Travail, the project involves two research institutions—CREST (Centre de Recherche en Économie et Statistiques) and LISN (Laboratoire Interdisciplinaire des Sciences du Numérique)—and brings together a multidisciplinary team of researchers: Guillaume Bied (University of Ghent), Philippe Caillou (UPSaclay - LISN - INRIA), Bruno Crépon (CREST), Christophe Gaillac (University of Geneva), Morgane Hoffmann (CREST), Solal Nathan (UPSaclay - LISN - INRIA), Mitia Oberti (CREST - France Travail) and Michèle Sebag (UPSaclay - LISN - INRIA - CNRS). The project has two main objectives. First, to develop a state-of-the-art job recommender system using machine learning to match job seekers with vacancies based on past hiring patterns. The second objective is to evaluate its impact on job seekers, employers, and the overall labor market through a series of field experiments. With access to the rich data of France Travail, which includes detailed information on job seekers, job vacancies, and

¹A CIFRE contract is an industrial agreement that funds PhD research in collaboration with a company or public institution.

²Valorisation des DONnées pour la Recherche d'Emploi.

³The project Vadore was supported by DATAIA convergence institute.

past matches, we successfully capitalized on artificial intelligence to develop a job RS matching job seekers with vacancies. This work has been recognized in publications at machine learning conferences [Bied et al., 2021b, 2023]. As part of the second objective, we have already obtained valuable insights from extensive offline empirical analyses and large-scale beta tests on job seekers (100,000 in March 2022, 170,000 in June 2023). The beta tests' results are currently being analyzed, and will be further enhanced by upcoming large-scale experiments designed to include repeated exposure to recommendations and assess spillover effects, contributing to a more comprehensive evaluation of the RS on the labor market.

Both Chapter 1 and Chapter 3 of this thesis directly stem from the Vadore project. Chapter 1, co-authored with the Vadore team, presents an offline study of different approaches to RS in the labor market, including the Vadore algorithm. Chapter 3, co-authored with Guillaume Bied, examines if job seekers have aversion towards algorithmic recommendations. The algorithms developed in the context of Vadore were employed in this study, and the findings provide valuable insights for the entire project. Chapter 2 of the thesis complements this work with a separate project co-authored with Abhijit Banerjee (MIT), Bruno Crépon (CREST-ENSAE), and Cécile Welter-Médée (INSEE). This chapter evaluates an increased personalization in the existing matching system employed by the French Public Employment Service.

The sequence of the chapters follows a logical progression. Chapter 1 lays the foundation by describing and comparing two representative types of job RS: (1) the Vadore algorithm, which is a state-of-the-art machine learning model predicting hiring chances, and (2) an expert system inspired by the proprietary algorithm of the French PES, based on job seekers' stated preferences. Through an extended offline comparative analysis and an economic model, it illustrates how each system captures distinct dimensions of the matching process, leading to different recommendations, and identifies an optimal recommendation system from the perspective of job seekers. One of the key challenges of RS design, highlighted in this chapter, is the measurement of the value of jobs for job seekers. This challenge is closely tied to Chapter 2, where we conduct an RCT to evaluate the impact of incorporating job seekers' stated preferences for job attributes into the recommendation system of the French PES. Rather than measuring job seekers' absolute valuation of each attribute, the focus is on assessing how they prioritize these attributes relative to one another to determine their overall utility from job ads. The analysis underscores both the benefits of personalization and the challenges involved in implementing it within real-world matching systems. Finally, in Chapter 3, we examine the phenomenon of algorithm aversion and highlight the heterogeneity in job seekers' attitudes towards algorithmic recommendations. This chapter offers a more behavioral perspective on how

users perceive and interact with algorithmic job RSs, complementing the previous chapters.

Together, these chapters open up numerous avenues for further research in the field of job RSs, both from an economic, algorithmic, and behavioral perspective. Much work remains to fully understand and optimize these technologies for the labor market.

Below, I describe each chapter in more detail.

Chapter 1: Designing Job Recommender Systems: How to Improve Human-Based Search

In recent years, there has been a surge in the use of recommendation algorithms in the labor market, reflecting the belief that the internet can enhance labor market efficiency and job matches' quality [Astor, 2001, Kuhn and Mansour, 2014b, Horton, 2017]. A variety of tools and approaches to generate such recommendations have been developed. Platforms and Public Employment Services (PESs), for example, have implemented tools to recommend job vacancies aligning with job seekers' preferences (see Broecke, 2023, Gutiérrez et al., 2019). Advanced Machine Learning (ML) tools have also emerged, proposing different types of recommendation systems, typically optimized using algorithmic criteria with limited consideration of individual behavior (RS hereafter, see, e.g., Freire and de Castro [2021] and proceedings of RecSys, which is a cycle of ML conferences exclusively focusing on issues related to RS). Indeed, while ML algorithms are adapted to different observational contexts, they primarily aim to predict the success of a match in various sense (predicting the hiring likelihood, of applying for a job, or even job seekers' interest).⁴ The underlying mechanism behind the potential efficiency gains of these algorithms is that they perform faster and at lower cost screening tasks that job seekers would do themselves.

At the same time, a number of studies have sought to measure and investigate the added value that RSs can bring. The existence of job search platforms and the possibility of large-scale experiments have also become a powerful means of studying search behavior and ways of improving it [Kircher, 2020, 2022]. While some studies are devoted to the recommendation of job seekers to recruiters (see for example Horton [2017] and

⁴Some algorithms, known as collaborative filtering strategies, rely solely on past choices to recommend jobs that other job seekers, who previously applied for the same jobs as the current user, have chosen (see, e.g., Le Barbanchon et al., 2023). Other algorithms, so-called *content-based recommenders* or *hybrid*, leverage comprehensive data about jobs, job seekers, and past choices [see, e.g., Volkovs et al., 2017, Zhao et al., 2021, Bied et al., 2021b, 2023, Mashayekhi et al., 2023]. They aim to predict matches by considering job seeker and vacancy characteristics. Some algorithms, instead, construct similarity measures based solely on the attributes of job seekers and job vacancies, without considering historical choices or matches.

Li et al. [2019]), many studies focus on recommendations to job seekers. Le Barbanchon et al. [2023] and Field et al. [2023] focus on recommendations of job vacancies to job seekers. Behaghel et al. [2024] considers firm recommendations to job seekers. An important area that has been studied is the recommendation of job seekers to broaden their search and consider other labor markets (Bélot et al. [2019, 2022], Altmann et al. [2022]). In this case, the underlying mechanism is that RS can redirect search efforts towards types of job vacancies where they would be more productive.

Several important questions emerge from the growing use of recommendation systems, particularly from the perspective of job seekers. A key issue is understanding the diversity of RS and identifying the conditions under which an RS effectively enhances a job seeker's outcomes. Central to this is the role of economic models in capturing the complexities of job search behavior. Another closely related challenge is determining whether there exists an optimal RS tailored to job seekers' needs, and if so, how such a system should be designed. This requires not only observing the choices made by job seekers and firms but also accurately modeling and estimating their underlying behavior. Finally, it is critical to assess whether simply replicating job seekers' past choices is a sufficient foundation for RS design, even when those choices are rational [see more broadly, e.g., Zhuang and Hadfield-Menell, 2020, Kleinberg and Raghavan, 2021, Kleinberg et al., 2022, Kasy, 2024, on value misalignment problem].

Other important questions also arise. First, if RSs improve job prospects by narrowing the set of relevant opportunities, they are likely to make job seekers more selective, as demonstrated by Kelley et al. [2024]. How does this increased selectivity influence the overall benefits derived from RSs? Should this behavioral change be factored into the design of these systems? Furthermore, it is essential to consider potential biases in job seekers' perceptions of their likelihood of success. As discussed by Bélot et al. [2019, 2022], Altmann et al. [2022], using RSs to redirect job search efforts to areas where they would be more productive is a promising strategy. However, how do these biases—along with the need to recommend vacancies that job seekers find appealing—constrain the design of effective RSs that truly benefit both job seekers and employers?

Leveraging large-scale administrative data that offer a uniquely comprehensive view of the search and matching process—encompassing the characteristics of job vacancies, job seekers' search parameters across various job dimensions, and their multiple interactions (e.g., clicks, applications, and hires)—this chapter aims to address these questions. We begin by highlighting the empirical significance of these issues, showing that different RSs capture distinct dimensions of the job search process and generate notably different recommendations. Specifically, we analyze two algorithms. The first, commonly used by PESs, matches job vacancies to job seekers based on declared search preferences. The

second, reflective of a growing body of machine learning literature, predicts successful matches by analyzing job seeker and vacancy characteristics. The rationale behind these designs, which we validate empirically, is that the first algorithm captures the utility \mathcal{U} a job seeker derives from a vacancy, while the second captures the recruitment chances \mathcal{P} of an application. We demonstrate that these two approaches yield highly divergent sets of recommendations in practice. To assess the potential of combining these approaches, we conducted a field experiment within the French PES. Job seekers were randomly assigned to receive recommendations from either the preference-based or ML-based algorithms, as well as from hybrid models that integrate both methods. The results show that while ML-based recommendations enhance perceived hiring chances, hybrid models that balance both hiring probabilities and job preferences generate higher engagement, particularly through increased click-through rates. These findings suggest that integrating both dimensions improves the relevance and effectiveness of job recommendations.

We develop a model to define the value of search when using a RS. The model incorporates two key factors: the utility of a job (U) and the chances of success of an application (P). This model helps address several of the previously raised questions. First, we show that there exists an optimal RS, and that optimal recommendations are based on a score that combines both U and P . Interestingly, while there may be ML-based algorithms to assess P (such as the one we employ, denoted by \mathcal{P}), no such algorithm exists to assess U . In fact, the version we use, denoted as \mathcal{U} , is derived from an expert system rather than an ML-based score. Second, combining these two factors, we demonstrate that the functional form of the optimal RS score cannot be learned solely by observing and replicating job seekers' behavior. The model also outlines the conditions under which an RS improves a job seeker's search value. Furthermore, the model shows that application behavior depends on the two previously mentioned factors, U and P , in a manner that closely aligns with the findings of [Hitsch et al. \[2010\]](#) in the context of online dating. By interpreting \mathcal{U} and \mathcal{P} as representations of these two key dimensions, we estimate the application model using job seekers' application data. Consistent with the model, we find that these two variables significantly influence the decision to apply for posted jobs. This constitutes the first empirical result of the chapter, validating the model and the two score representations, and providing evidence that both are central to job seekers' decision-making processes.

The estimates derived above allow us to reconstruct the valuation functions for specific vacancies for each job seeker, as well as to determine the optimal recommendation system (RS). Using these results, we evaluate the recommendation sets produced by the two initial RSs—based on the \mathcal{U} and \mathcal{P} scores—and the optimal algorithm. By comparing these values to a benchmark that reflects job seekers' own searches, while holding screen-

ing effort constant, we are able to quantify the value added by each RS. This forms the second key empirical contribution of the chapter.

Our findings reveal that the optimal RS and the ML-based RS (driven by \mathcal{P}) perform similarly, both adding substantial value to the job search process for job seekers. In contrast, the algorithm relying on matching search parameters (i.e., \mathcal{U}) performs significantly worse. In fact, it leads to a decrease in the value of job seekers' search efforts when compared to the benchmark.

Importantly, we find that the performance of the optimal RS varies widely across individuals. While the optimal RS substantially improves outcomes for job seekers with strong labor market prospects, the benefits are far more modest for those with weaker prospects. This heterogeneity underscores the fact that RSs, while adding value overall, may exacerbate existing labor market inequalities by disproportionately benefiting those already well-positioned.

Overall, our study underlines the importance of identifying several essential quantities for building a high performance RS, in particular the job utility (U) and the probability of a successful application (P). The latter can be derived from the available data, a task at which ML tools excel, while exploiting the complexity of the information available on both sides of the market. However, identifying a prediction of the job seekers' utility associated with a position, seemingly simpler, is challenging due to the lack of direct observation.

In essence, this chapter relates different strands of the literature. Given the importance of online job search [Kuhn and Mansour, 2014b, Kircher, 2022], the first is related to the impact of recommendations on labor market frictions. Following B elot et al. [2019], several studies show that suggestions to expand the search to alternative occupations have an effect on interviews and future job outcomes [B elot et al., 2022, Altmann et al., 2022]. Relatedly, some studies discuss how the identification of sectors to be recommended can be refined by collecting specific data on skills from job seekers [B achli et al., 2024]. While in the previous papers, broadening job search to recommended job markets is a suggestion, sometimes taking the time of directly showing vacancies as in B elot et al. [2019], broadening job search can be made compulsory. van der Klaauw and Vethaak [2022] show that in such a case, the impact can be negative on both job search and job quality, which might suggest heterogeneous impact and room for improving recommendations.⁵ Other studies instead recommend firms predicted to hire in a broaden labor market: Behaghel et al. [2024] show in a randomized experiment in which job seekers are randomly assigned to treatment as well as the firms recommended to them, that being assigned

⁵In this study, broadening the job search means applying for jobs in other occupations or geographical areas, offering a lower salary or requiring lower skills.

to receive recommendations has an impact, although small, on hiring, mainly at recommended firms. Drawing on the specifics of their experimental design, they show that recommended firms indeed and as expected have hiring potential conditional on application, and that the treatment's effect on job seeker-firm matching can be attributed to an increase in the application rate on recommended pairs. Other studies instead use job seekers' interest in certain vacancies, using click history to generate recommendations: [Le Barbanchon et al. \[2023\]](#) shows in a two-sided experiment that such RS redirects job seekers' search and hiring towards recommended vacancies, which themselves receive more applications. As these studies show, and sometimes discuss [[Behaghel et al., 2024](#)], there are several conditions for these RS to be effective. They need to arouse the interest of job seekers, but also direct them towards vacancies with good recruitment potential. And as they show, it is not out of the question that recommendations may lead to a deterioration in the situation of job seekers. It is worth to note that these studies use very different strategies to generate recommendations. Our contribution is to illustrate firstly that recommendations based on different principles do, in fact and quite logically, generate highly different recommendations, both in terms of chances of recruitment and their attractiveness from the job seeker's point of view. Our contribution is then, thanks to a data-validated theoretical model, to identify the conditions necessary for a RS to improve the situation of job seekers, and to define one RS that is optimal from their point of view. One key insight of our analysis is that RS often focus on a precise objective - typically improving the chances of a match - which may differ from the job seekers' objectives. This disconnection between the two can result in substantial losses, as some individuals may focus their search on vacancies far from their preferences following the recommendations.

In addition, the literature highlights the importance of taking into account the behavioral aspects of the job search [see, e.g., [Babcock et al., 2012](#), [Cooper and Kuhn, 2020](#), [Altmann et al., 2018](#)], whether they be, among others, biased perceptions of the chances of success of search strategies or biased perceptions of market fundamentals. Indeed, recent empirical work has highlighted the existence of a category of job seekers who remain permanently over-optimistic about their chances of returning to work [[Mueller et al., 2021](#), [Mueller and Spinnewijn, 2023](#)]. These biases could prevent job seekers from applying for high-yield vacancies: [Field et al. \[2023\]](#) show in an experiment involving recommendations made to job seekers, that the inertia of behavior due to the simple fact of calling a firm to apply for a potentially interesting recommended vacancy can be a major barrier to applications. They also show that lowering this psychological cost greatly increases the number of applications. More directly, as shown by [Kelley et al. \[2024\]](#), being exposed to a better pool of vacancies may lead job seekers, as predicted by the basic search model, to adjust their search behavior by being more selective about which vacancies to apply for.

Our contribution is, within the framework of a sufficiently flexible model, to discuss how behavioral adjustment or the existence of behavioral biases can affect the effectiveness of RS and how they could be taken into account in its design. This chapter highlights the importance of designing RS based on both individual hiring prediction and a relevant and detailed representation of behaviors, starting for example, with the importance of accurately measuring preferences for different job attributes [Mas and Pallais, 2017a, Wiswall and Zafar, 2018, Feld et al., 2022a, Banerjee et al., 2022, Banerjee and Chiplunkar, 2024].

Focusing on this central aspect of basing RS on job seekers' behaviors, this chapter leaves in the shade the important dimension of labor market congestion and the impact of RS on congestion. Yet congestion is a key dimension of the functioning of RS. It is central to several of the papers cited above [Kircher, 2022, Behaghel et al., 2024, Le Barbanchon et al., 2023] which, despite large-scale randomized experiments, sometimes fail to provide evidence on such congestion effects. Congestion is also instrumental in our previous work Bied et al. [2021b,a], which sketches the path of optimal transport as a means of dealing with this problem. In this contribution, we focus on the instrumental dimension of behavior to design effective RS, with the feeling that potential issues that congestion may pose when designing RS will be more easily and effectively studied once an effective recommender score has been developed.

Chapter 2: Personalization Pitfalls: The Unintended Effects of Using Stated Preferences Data in Job Recommendations

Addressing the multifaceted nature of unemployment requires a close examination of the persistent matching frictions within the labor market. These frictions create a paradoxical situation where job vacancies coexist with an available workforce, leading to prolonged unemployment spells and inefficient utilization of human resources. Such frictions arise from informational asymmetries, geographic mismatches, and the complexities of aligning workers' skills with job requirements.

As a response to these challenges, Public Employment Services (PES) have turned to algorithmic solutions to better match job seekers with vacancies. Research has shown that automated job recommendations can lead to improvements in interview rates [Bélot et al., 2018, Li et al., 2020] and long-term employment outcomes [Bélot et al., 2022, Altmann et al., 2022, Behaghel et al., 2024], although some studies report only modest effects [Le Barbanchon et al., 2023]. Among the different types of job recommendation algorithms used, expert systems or "knowledge-based" systems are a notable approach, accounting for about 12.7% of job recommender systems in recent studies [Freire and

de Castro, 2021] and are widely implemented in PES globally [World Bank, 2023]. Unlike machine learning models that rely on data-driven inferences, expert systems operate based on explicit, predefined rules crafted by human experts. These “if-then” rules consider a multitude of factors related to both job seekers and job vacancies, offering a structured way to address mismatches in the labor market.

An illustrative example of such an expert system is the ELISE Search and Match platform developed by WCC, a Dutch technology company. The ELISE platform is utilized by various European countries—including France, Luxembourg, Austria, and Germany—in their public employment services to match job seekers with employers [World Bank, 2023]. In France, this matching system plays a central role in the operations of the PES. It is integrated into the PES online platform used by job seekers, serving both as a search engine and a recommender system. Recruiters also use the same platform to find suitable candidates for their vacancies. Additionally, PES caseworkers utilize this matching algorithm to identify suitable candidates for positions where employers have specifically requested assistance in finding appropriate candidates.

The matching process begins by collecting detailed information from job seekers, such as their qualifications, work experience, skills, and specific search criteria—including desired occupation, reservation wage, preferred location, contract type, and working hours. Job vacancies are similarly detailed, encompassing factors like job requirements, geographical location, salary, contract type, and working hours. The expert system applies its predefined rules to assess the compatibility between a job seeker’s profile and each job vacancy. For instance, one rule might verify if the offered salary meets the job seeker’s reservation wage, while another evaluates whether the job’s location is within a reasonable commuting distance for the job seeker. To balance the various factors involved in matching, each criterion is assigned a weight that reflects its importance in the overall process. These weights are determined by human experts who leverage their understanding of labor market dynamics to prioritize certain attributes over others. By calculating a matching score for every job seeker–vacancy pair—where higher scores indicate better matches—the system identifies the most suitable job opportunities for each individual. The vacancies with the highest matching scores are then recommended to job seekers, aiming to efficiently connect them with positions that best meet their needs and qualifications.

Despite these advancements, a significant limitation remains in the uniform application of attribute weights, leading to a lack of personalization. The French PES matching system applies the same weights to each attribute across all job seekers, failing to account for the varied priorities of individuals. Indeed, several studies highlight heterogeneity in workers’ preferences for job attributes. Workers value various job characteristics differ-

ently, and these preferences can significantly influence job choices and contribute to labor market outcomes. For instance, [Wiswall and Zafar \[2017\]](#) explore how gender differences in workplace preferences affect job choices, human capital investments, and contribute to the gender wage gap. Similarly, [Mas and Pallais \[2017b\]](#) examine workers' valuation of flexible scheduling and work-from-home options, finding that women, especially those with children, have a higher willingness to pay for these attributes. Using French administrative data [Le Barbanchon et al. \[2021\]](#) finds that unemployed women value shorter commutes around 20% more than men, resulting in a willingness to accept lower wages. [Maestas et al. \[2023\]](#) investigate workers' willingness to trade wages for better job characteristics, finding that preferences vary across demographic groups and contribute to wage inequality. These findings underscore the importance of incorporating individual preferences into job matching algorithms to better cater to diverse job seeker priorities. Additionally, [Banerjee and Chiplunkar \[2024\]](#) found that Indian job seekers prioritize salary, location, and job title the most. Notably, the study reveals that job location holds greater importance for women than men, reflecting a preference for jobs closer to their homes.

Incorporating user input into these systems could address the issue of personalization. For instance, [Banerjee and Chiplunkar \[2024\]](#) showed that a more tailored matching system, which considers diverse priorities and preferences of job seekers, could potentially enhance job seeker satisfaction and employment outcomes. For quite some time, researchers in computer science have recognized that the effectiveness of recommender systems extends beyond mere accuracy [[Swearingen and Sinha, 2001](#)], underscoring the significance of user involvement in the recommendation process. Incorporating individual preferences and feedback leads to more precise and relevant suggestions, enhancing not only the system's accuracy but also its relevancy [[Pu et al., 2012](#), [Parra and Brusilovsky, 2015](#)]. For instance, systems like "TasteWeights" enable users to fine-tune weights on different parameters, which not only improves the accuracy of music recommendations but also significantly enhances user engagement and satisfaction [[Bostandjiev et al., 2012](#)]. In the context of job search, [Bächli et al. \[2024\]](#) propose a method that allows job seekers to fine-tune the parameters of an occupational recommender system. These parameters (α and β) enable job seekers to either target their own skill profile or previous occupation (α), and anchor suggestions to other requirements in their previous occupation (β).

However, involving users in the recommendation process introduces a trade-off between the recommendations' accuracy and user effort. While eliciting detailed preferences can enhance recommendation precision, it may also increase cognitive load and user burden, potentially reducing user engagement [[Pu et al., 2012](#)]. One explanation from behavioral decision theories is that individuals tend to prioritize minimizing cognitive effort over maximizing accuracy because the effort is immediately perceptible,

whereas the benefits of increased accuracy are delayed and uncertain [Häubl and Trifts, 2000]. Furthermore, studies have shown that while collecting user input initially enhances recommendation precision, there are diminishing returns to additional user effort, with the most substantial gains in recommendation quality often achieved with only minimal input (e.g. Drenner et al. [2008]). This underscores the importance of balancing accuracy with user effort to maintain user engagement.

Therefore, selecting an appropriate method for eliciting user preferences is essential to balance personalization accuracy and user engagement. Preference elicitation methods are generally classified into revealed and stated preference approaches. Revealed preferences infer choices from observed behaviors but are limited to existing options and cannot assess hypothetical scenarios. Stated preferences directly solicit individual preferences, allowing for hypothetical contexts but may face concerns about hypothetical bias and external validity [Diamond and Hausman, 1994, Manski et al., 2000]. Interestingly, Banerjee and Chiplunkar [2024] found that job seekers' reported preferences regarding job aspirations and priorities were consistent regardless of whether incentives were provided to elicit their *true* preferences. This consistency suggests that the preferences elicited were genuine rather than strategic responses aimed at increasing their chances of securing a job.

Within stated preference methods, there are two main approaches: compositional and decompositional [Helm et al., 2004, Weernink et al., 2014, Marsh et al., 2016]. Compositional methods directly ask individuals to allocate importance to different job attributes, while decompositional methods, such as discrete choice experiments (DCE), infer preferences based on the choices individuals make between different job options. While DCE has gained popularity among labor economists as an alternative to revealed preference methods for estimating compensating wage differentials [Mas and Pallais, 2017a, Wiswall and Zafar, 2017], our study employs a compositional approach to capture job seekers' preferences more directly. Common compositional techniques include direct rating, direct ranking, and point allocation (for a review, see Van Ittersum et al. [2007], Zardari et al. [2014]). However, direct rating and ranking methods often lack differentiation among attributes due to the absence of trade-offs, leading participants to rate all attributes as highly important [Krosnick and Alwin, 1988, McCarty and Shrum, 2000]. The point allocation method addresses this issue by requiring individuals to distribute a fixed number of points among attributes, introducing explicit trade-offs and encouraging careful consideration of attribute importance. Although the direct-rating method shows higher test-retest reliability due to its simplicity [Bottomley et al., 2000], this does not necessarily indicate greater validity in capturing true preferences [Bottomley and Doyle, 2013].

Given our objective to derive precise attribute weights for personalizing job recom-

mentations, we use the point-allocation method in our study. This approach effectively engages respondents in trade-offs that reflect their true preferences without imposing excessive cognitive burdens. We allocate 15 points to participants, asking them to distribute these among five job attributes: occupation, commuting distance, wage, contract type, and working hours. This method allows job seekers to indicate the relative importance of each attribute, providing data that can be directly applied as weights in our personalized job matching algorithm.

With this perspective, our research aims to explore the impact of integrating these individual job seekers' preferences into the PES matching algorithm on recommendation appreciation and adoption. We also aim to understand the variations in these preferences according to socio-demographic factors and unemployment history. To this end, we have implemented an intervention involving a personalized version of the PES matching algorithm that recommends job vacancies to job seekers based on their preferences regarding the importance of the following job attributes: occupation, wage, distance from home, working hours, and type of contract. This intervention is facilitated through a web interface designed to collect job seekers' preferences about job attribute importance and subsequently offer them job vacancies that align with these personalized preferences. Our study is divided into two sequential phases. In the first phase, we conduct a pilot study to test and verify the reliability of the point-allocation method for collecting job attribute preferences. Participants in this phase are asked to distribute a fixed number of points among key job attributes (occupation, salary, commuting distance, type of contract, working hours) to capture the relative importance of each attribute. In the second phase, a large-scale randomized controlled trial (RCT) is implemented, where participants state their preferences using the same method as in the pilot study. The RCT involves three arms: one third of job seekers (henceforth the "full treatment" group) receiving recommendations tailored to their personalized weights, one third ("partial treatment" group) providing preferences but receiving standard algorithm recommendations, and one third (control group) following the existing PES algorithm without preference input. This design allows us to assess the impact of preference-based personalization on job seeker engagement. The pilot experiment was conducted on approximately 20,000 job seekers in March 2022, before launching a full-scale randomized controlled trial on 250,000 job seekers in June, 2022.

The pilot study revealed that job seekers generally engaged thoughtfully with the preference elicitation tasks, indicating the reliability of the point-allocation method used to capture job attribute preferences. In the large-scale experiment, job seekers' responses to the point allocation task showed a considerable divergence from the standard weights used in the PES matching algorithm, with preferences displaying a more balanced distri-

bution across attributes. Notably, wage and commuting distance were rated higher than occupation, which contrasts with the emphasis placed on occupation by the standard PES system.

Our findings reveal that integrating personalized preferences into the recommender system resulted in an increased number of job recommendations. This occurred because, with the standard weights used in the PES algorithm, many job ads did not accumulate sufficient scores to be recommended to job seekers. By incorporating job seekers' own point allocations, more ads met the criteria for recommendation. Consequently, while the number of recommendations increased, some of these ads included positions that were less aligned with job seekers' preferred occupations and salary expectations.

Then, we examine interactions between job seekers and the job recommendations made during the experiment (clicks, intentions to apply and actual applications). We do not look at organic clicks and applications on the PES platform. To isolate the effect of personalization, we compared job seekers in the full treatment group (receiving personalized recommendations) with those in the partial treatment group (receiving standard recommendations). We find that personalization alone showed significant and positive effects : job seekers in the full treatment group clicked on recommendations 13.8% more, exhibited a 20.8% increase in intentions to apply, and a 23.1% rise in actual applications compared to the partial treatment group. These effects were primarily driven by the new job ads introduced through the personalized algorithm rather than a mere reallocation of interest from standard recommendations.

However, these positive outcomes were not enough to counterbalance the effort expenditure made by treated individuals when completing the preference elicitation task. When comparing the full treatment group to the control group, we found a 26% reduction in clicks, a 27% decline in intentions to apply, and a 30% drop in actual applications. Additionally, platform visits decreased by 24%, indicating reduced overall engagement. Further, we examined the impact of the point allocation task itself by comparing the partial treatment group to the control group. Here, we observed a sharp decline in engagement: clicks decreased by 35%, intentions to apply fell by 40%, and applications dropped by 43%. Platform visits also declined by 25.7%. This suggests that the point allocation task set higher expectations for job recommendation accuracy. When these heightened expectations were not met, job seekers experienced disappointment, leading to reduced engagement. Overall, our results align with previous research on the trade-off between accuracy and effort in recommender systems [Pu et al., 2012]. They highlight that the benefits of personalization must be weighed against the potential for increased user burden, emphasizing the need for careful design in the implementation of preference-based algorithms.

The findings have important implications for the design of job recommender systems. Personalized recommendations lead to better engagement; hence, policymakers and platform designers might consider incorporating more user-driven features into these systems. However, the benefits of personalization must be weighed against the potential for increased user burden. Balancing personalization with user effort seems essential to maximize the effectiveness of recommender systems in the labor market.

This chapter contributes to several streams of literature. First, we contribute to the research on the impact of recommender systems on the labor market by providing empirical evidence on how personalized job recommendations affect job seeker engagement and recommendation adoption within a PES context. Our findings add to the evidence on the effectiveness of job recommender systems [Bélot et al., 2018, Li et al., 2020, Bélot et al., 2022, Altmann et al., 2022, Kuhn and Mansour, 2014a, Le Barbanchon et al., 2023, Behaghel et al., 2024]. Second, we contribute to the literature on job attribute preferences by analyzing how job seekers' preferences vary according to socio-demographic factors and unemployment history, aligning with previous studies that highlight heterogeneity in workers' valuation of job attributes [Wiswall and Zafar, 2017, Mas and Pallais, 2017b, Le Barbanchon et al., 2021, Feld et al., 2022b, Maestas et al., 2023]. Third, we contribute to the literature on different types of recommender systems in job matching by evaluating a personalized expert system that incorporates user input [Freire and de Castro, 2021].

Chapter 3: Job Seekers' Responses to AI Job Recommendations: Insights from a Field Experiment

Recommender systems (RS) have become a ubiquitous feature of online platforms [Aggarwal, 2016], influencing numerous aspects of our daily lives by suggesting items tailored to our individual preferences (e.g., songs or movies on streaming services, products on e-commerce sites, and content on social media feeds). These systems address the issue of information overload by filtering vast amounts of data, providing users with relevant options, and enhancing their overall experience. In the context of job search, job recommender systems, a specific type of RS, employ AI models to match job seekers with vacancies that align with their skills, preferences, and career goals [Mashayekhi et al., 2024]. By analyzing the extensive data from online job platforms, including user behavior and vacancy characteristics, these systems aim to reduce search costs and improve job matching in the labor market.

Recognizing this, Public Employment Services (PES) in many countries are adopting AI-driven RSs to address inefficiencies in traditional job-matching processes, provid-

ing personalized, data-driven recommendations tailored to individual job seekers [World Bank, 2023]. Field experiments, such as those conducted by B elot et al. [2019, 2022], Altmann et al. [2022], show that automated occupation recommendations can expand job seekers’ application scope, increase their chances of securing interviews, and improve employment metrics like hours worked and earnings. Similarly, studies focusing on personalized recommendations for companies likely to recruit in France [Behaghel et al., 2024] and on personalized job recommendations in Sweden [Le Barbanchon et al., 2023] demonstrate positive, albeit modest, effects on employment rates.

While promising, these outcomes are constrained by a considerable challenge – algorithm aversion. This phenomenon, in which users are reluctant to trust or engage with algorithmic recommendations to the same extent as they would with human-generated ones, restricts the comprehensive integration of AI tools in job search and related markets. To fully capitalize on the potential of AI-driven job recommender systems, it is essential to comprehend and address this aversion.

The literature on attitudes towards algorithms offers two competing perspectives. The concept of algorithm aversion was first introduced by Dietvorst et al. [2015], who identified that users tend to lose trust in algorithms more quickly than in humans, particularly following errors. Subsequent studies, such as those by Jussupow et al. [2020] and Mahmud et al. [2022], expanded the definition of algorithm aversion into a broader preference for human decision-making, even when no errors are present. In particular, Jussupow et al. [2020] define algorithm aversion as a biased evaluation process in which users undervalue algorithmic decisions compared to those made by human agents, who can range from recognized experts to average individuals. This suggests that reducing the visibility of recommendations’ algorithmic origins in job recommendations could mitigate aversion. Conversely, another stream of research points to algorithm appreciation, where users exhibit trust and even preference for algorithmic recommendations, especially when algorithms are perceived as objective, data-driven, and suitable for the task at hand [Jussupow et al., 2020, Mahmud et al., 2022]. This perspective implies that making users aware that recommendations are algorithmically generated could enhance their engagement with the system.

This leads us to the central question: how should algorithmic recommendations be framed to foster user engagement? To address this question, we conduct a large-scale Randomized Controlled Trial (RCT) with the French PES, employing a between-subject design to test job seekers’ responses to recommendations labeled as algorithmic, human-curated, or without any source attribution. In our experiment, job seekers receive recommendations generated by a single hybrid model that integrates an expert system module—simulating human decision-making—with a state-of-the-art machine learning (ML)

model trained on past hiring data (described in [Bied et al. \[2023\]](#)). This setup allows us to randomly vary the displayed source, presenting the recommendations as either human-driven (expert system) or algorithmically-driven (ML model) without changing their content. Participants are presented with five job ads recommended by our hybrid algorithm through an online survey designed to mimic real-world job search experiences. They are first asked to rate each recommendation using “thumbs up/thumbs down” buttons. Next, the same ads are displayed with clickable links, allowing participants to visit the PES page for each job and apply if they wish. We measure job seekers’ aversion by comparing the outcomes of the algorithm group to those of the human group, in line with the literature. We focus on three key outcomes, capturing the engagement with recommended ads at different stages: initial interest, clicks for more details, and application submissions.

Our analysis reveals a clear aversion to algorithmic job recommendations. On average, job seekers exposed to algorithm-labeled recommendations express significantly less interest and engage less with the recommendations, as indicated by a lower number of clicks. We estimate that algorithm aversion reduces the total number of declared interests (“thumbs up”) and clicks by approximately 10% and 15% respectively, compared to recommendations framed as coming from human experts. While these effects are particularly evident in early-stage engagement metrics like interest and clicks, we do not observe statistically significant effects on job applications, likely due to limited statistical power at this final stage of engagement.

We further explore the role of recommendation quality in shaping algorithm aversion. Quality in recommendation systems can be assessed in various ways, such as the predictive accuracy of recommended matches (i.e., whether the recommended matches will lead to actual hires) or the alignment with user preferences. We choose the latter approach because it is directly observable to job seekers, serving as a practical proxy for perceived quality from the user’s perspective. Specifically, our quality measure is a job seeker-centered matching score that evaluates how well a recommendation aligns with the job seeker’s stated preferences across multiple dimensions: occupation, contract type, location, salary, and working hours. Our findings reveal that high-quality recommendations—whether labeled as algorithmic or human-generated— elicit more engagement from job seekers. However, the aversion to algorithm-labeled recommendations persists regardless of the recommendation quality. In other words, even when the algorithm provides matches that closely align with user preferences, users remain less inclined to engage compared to when recommendations are framed as human-generated. This underscores the importance of understanding how individual job seekers characteristics influence responses to algorithmic systems and indicates that addressing algorithm aversion

requires more than just improving recommendation quality.

Following the idea that algorithm aversion may be influenced more by user characteristics than by recommendation quality, we conduct a more detailed analysis of this phenomenon. Using machine learning techniques, specifically the approach developed by Chernozhukov et al. [2023], we identify varying responses among different subgroups of job seekers to algorithm-labeled recommendations. Our analysis reveals that attitudes towards algorithmic recommendations are diverse. Interestingly, some job seekers show a preference for algorithmic recommendations over those presented as human-curated, a behavior known as “algorithm appreciation” [Logg et al., 2019]. We identify two main groups: the 20% most “algorithm-friendly” job seekers, who respond most positively to algorithm-labeled recommendations, and the 20% most “algorithm-averse” job seekers, who show the strongest preference for recommendations framed as human-generated. For the algorithm-friendly group, the ATE of being informed that the recommendations are algorithmic rather than human is positive, leading to a 71% increase in the interest rate per job recommendation and a 46% increase in the click rate, relative to the means for the human-labeled group (0.18 for interest rate and 0.07 for click rate). In contrast, the algorithm-averse group shows a negative ATE: being told that the recommendations are algorithmic results in an approximately 103% decrease in the interest rate and a 79% decrease in the click rate compared to the human-labeled group means. On average, there are more algorithm-averse job seekers, which accounts for the overall aversion observed in our study.

By focusing on these two extreme groups, we are able to identify distinct patterns that help to elucidate the characteristics that contribute to these opposing attitudes. In this analysis, we utilise clicks on recommendations as the primary outcome measure, as this is the most policy-relevant indicator of job seeker engagement, directly reflecting the likelihood of further exploration of job opportunities. The results indicate that younger job seekers, particularly those under the age of 35, are disproportionately represented within the algorithm-friendly group. Furthermore, this group is also characterised by a higher proportion of job seekers who are facing prolonged periods of unemployment and approaching the exhaustion of their unemployment benefits. This suggests that these individuals may have a more pressing need for effective job placements. Additionally, those who are less autonomous and who receive intensive support from the PES (who see their case worker more often) are over-represented in this group. Moreover, they are disproportionately represented in labor markets where the PES platform has a high level of market penetration. This indicates that a significant proportion of job vacancies are posted on the PES platform, in comparison to other general job platforms that are not specifically designed for registered job seekers. Furthermore, they are over-represented

in tighter labor markets, where the number of job vacancies exceeds the number of job seekers. Conversely, older job seekers, particularly those over the age of 50, are more prevalent in the algorithm-averse group. This group also exhibits higher wage expectations and shorter periods of unemployment. The PES provides only moderate to minimal support to this group, and they are more likely to be in labor markets with lower PES market penetration, where relatively few job ads are posted on the PES platform in favour of other general job platforms. Additionally, they are more commonly found in competitive labor markets, where the number of job seekers surpasses available vacancies. These findings indicate that both algorithm aversion and appreciation are shaped by a combination of factors, including age, unemployment duration, labor market conditions, and the level of support provided by the PES.

From a policy perspective, our findings indicate that enhancing the quality of recommendations may not be a sufficient strategy for mitigating algorithm aversion. Instead, it is imperative that recommendations are carefully framed in order to minimize the negative effects of algorithm labeling and to maximize the benefits of these tools. Our results also demonstrate that there is no universal approach to framing recommendations. The heterogeneity in job seekers' responses indicates that the framing of the recommendations should be tailored based on individual characteristics, or alternatively, that a neutral framing should be used that allows job seekers to opt into learning more about the recommendation source if desired.

Our study contributes to several strands of the literature. First, a key contribution is our focus on algorithm aversion within the context of a high-stakes task: job recommendations for job seekers registered at a Public Employment Service (PES). Although algorithm aversion has been the subject of investigation in domains such as joke recommendations and financial forecasting, the implications are considerably more significant in the context of the labor market. This is underscored by the European Union's AI Act, which categorizes AI systems utilized in employment as "high-risk" applications [European Commission, 2021]. The AI Act acknowledges the potential for AI-driven decisions in recruitment and job placement to have significant and enduring consequences for individuals. Indeed, research indicates that user responses to algorithms are contingent upon the nature of the task. In the context of subjective tasks, such as entertainment or tasks involving personal preferences, users tend to prefer human judgment. This is based on the assumption that algorithms are unable to adequately capture individual uniqueness or personal context [Prah and Van Swol, 2017, Yeomans et al., 2019, Longoni et al., 2019]. Conversely, algorithms are seen as more effective for objective, data-driven tasks where human-like decision-making is less necessary [Castelo et al., 2019]. In the HR field, perceptions of algorithmic versus human recruiters vary, with studies report-

ing mixed results regarding their perceived fairness, competence, trust, and usefulness [van Esch et al., 2021, Wesche and Sonderegger, 2021, Choung et al., 2023]. In the context of job ad recommendations, Ochmann et al. [2020] found that integrating human-like characteristics into AI job recommender systems can enhance their acceptance among job seekers. A closely related study by Bana and Boudreau [2023] investigated algorithm aversion in a university-based labor market platform, focusing on undergraduate and graduate students. Their study revealed that job seekers were less likely to pursue opportunities recommended by an algorithm compared to those recommended by a human manager, highlighting significant algorithm aversion. However, their context—students on an academic job platform—differs substantially from that of our study. The job seekers registered with the French PES represent a more diverse and vulnerable population, facing a wider array of employment challenges. Thus, while Bana and Boudreau [2023] provides valuable insights, our study extends the literature by focusing on a broader and more representative sample, exploring algorithm aversion in a real-world setting where engagement with algorithmic tools directly impacts employment outcomes.

Our second contribution lies in exploring how the quality of job recommendations—specifically, the alignment between algorithmic recommendations and job seekers’ preferences—affects algorithm aversion. Existing literature suggests that an algorithm’s performance significantly shapes user attitudes, as users often lose trust in algorithms after encountering errors, emphasizing the importance of accuracy in mitigating algorithm aversion [Dietvorst et al., 2015, Bogert et al., 2021]. However, in job recommendation systems, users typically lack visibility into the algorithm’s absolute accuracy, which is defined as its ability to match job seekers with positions that result in actual hires. In this context, Laumer et al. [2018] identified “performance expectancy” (i.e., the perceived usefulness of a technology [Venkatesh et al., 2003]) as a critical factor influencing job seekers’ intentions to use these systems. Yet, their study did not investigate how job seekers’ reactions differ between algorithmic and human-generated recommendations, leaving a gap in understanding how perceived performance affects algorithm aversion. Furthermore, the literature highlights the importance of “outcome favorability”—whether the results produced by an algorithm are seen as favorable or beneficial to the user—as a key factor in user attitudes. The “outcome favorability bias” refers to the tendency of users to view a decision-making process more positively when they receive favorable outcomes [Wang et al., 2020]. In Choung et al. [2023]’s study, a 2x2 design was used in which both the source of the evaluation (AI algorithm or human) and its favorability (favorable or unfavorable) were randomized across participants. After reviewing a job candidate’s profile for a job opening, participants were shown an evaluation of this profile, either attributed to an AI or a human. The study found that outcome favorability bias was stronger when

the evaluation was attributed to a human, i.e. participants reacted more negatively to unfavorable outcomes from a human than from an AI. However, how outcome favorability (i.e., the alignment between users' expectations and the recommended items) affects advisory algorithms, such as job recommendation systems, remains under-explored.

The third contribution of this chapter is a systematic exploration of how user characteristics shape attitudes toward algorithmic job recommendations. Using a data-driven approach free from a priori assumptions, we uncover strong heterogeneity in job seekers' responses, revealing the simultaneous existence of both algorithm-averse and algorithm-friendly groups. This extends previous research, which has demonstrated that factors such as psychological traits, demographics, and familiarity with algorithms significantly influence attitudes toward algorithmic systems, as detailed in [Mahmud et al. \[2022\]](#). For instance, individuals with high self-esteem may feel demeaned by algorithmic evaluations, while those confident in their abilities often prefer their own judgment over algorithmic suggestions, especially in areas where they have expertise [[Lee, 2018](#), [Logg et al., 2019](#)]. Attitudes also vary by age and gender: older individuals and women, in some cases, perceive algorithmic decisions as less beneficial, though this perception is not always consistent across sectors [[Araujo et al., 2020](#), [Logg et al., 2019](#), [Thurman et al., 2019](#)]. Additionally, those with lower education levels and less comfort with numbers tend to have a lower appreciation of algorithms across different industries [[Logg et al., 2019](#)]. In the HR field, [Pethig and Kroenung \[2023\]](#) found that women are more likely to choose algorithmic over human evaluators across different hiring and career-development settings, particularly when the human evaluator is male, due to the perceived objectivity and fairness of algorithms in mitigating gender biases.

Finally, unlike previous studies that have predominantly relied on vignette or lab environments, our approach captures the natural behaviors of job seekers interacting with real-world job recommendations within the French PES. By analyzing how algorithm aversion manifests in this public employment setting we are able to provide a more comprehensive understanding of the factors influencing the adoption and effectiveness of job recommender systems.

Introduction References

- Charu Aggarwal. Recommender Systems, volume 1. Springer, 2016.
- Steffen Altmann, Armin Falk, Simon Jäger, and Florian Zimmermann. Learning about job search: A field experiment with job seekers in germany. Journal of Public Economics, 164:33–49, 2018.
- Steffen Altmann, Anita M Glenny, Robert Mahlstedt, and Alexander Sebald. The direct and indirect effects of online job search advice. 2022.
- Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. In ai we trust? perceptions about automated decision-making by artificial intelligence. AI & society, 35:611–623, 2020.
- David Autor. Wiring the labor market. Journal of Economic Perspectives, 15(1):25–40, 2001.
- Linda Babcock, William J Congdon, Lawrence F Katz, and Sendhil Mullainathan. Notes on behavioral economics and labor market policy. IZA Journal of Labor Policy, 1:1–14, 2012.
- Mirjam Bächli, Hélène Benghalem, Doriana Tinello, Damaris Aschwanden, Sascha Zuber, Matthias Kliegel, Michele Pellizzari, and Rafael Lalive. Ranking occupations by their proximity to workers’ profiles. Swiss Journal of Economics and Statistics, 160(1):8, 2024.
- Sarah Bana and Kevin Boudreau. Behavioral responses to algorithmic matching: Experimental evidence from an online platform. 2023.
- Abhijit Banerjee, Bruno Crépon, Elia Pérennès, and Cécile Walter-Médée. Using stated individual preferences for job ads’ attributes to design a better matching algorithm between job ads and job-seekers. AEA RCT Registry, 2022. doi: 10.1257/rct.8719.
- Abhijit V Banerjee and Gaurav Chiplunkar. How important are matching frictions in the labor market? experimental & non-experimental evidence from a large indian firm. Journal of Development Economics, page 103330, 2024.
- Luc Behaghel, Sofia Dromundo, Marc Gurgand, Yagan Hazard, and Thomas Zuber. The potential of recommender systems for directing job search: A large-scale experiment. 2024.

- Guillaume Bied, Philippe Caillou, Bruno Crépon, Christophe Gaillac, Victor Alfonso Naya, Elia Pérennes, and Michèle Sebag. Designing labor market recommender systems: the importance of job seeker preferences and competition. In 4. IDSC of IZA Workshop: Matching Workers and Jobs Online-New Developments and Opportunities for Social Science and Practice, 2021a.
- Guillaume Bied, Elia Pérennès, Victor Alfonso Naya, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michele Sebag. Congestion-avoiding job recommendation with optimal transport. In FEAST workshop ECML-PKDD 2021, 2021b.
- Guillaume Bied, Solal Nathan, Elia Perennes, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michèle Sebag. Toward job recommendation for all. In Thirty-Second International Joint Conference on Artificial Intelligence {IJCAI-23}, pages 5906–5914. International Joint Conferences on Artificial Intelligence Organization, 2023.
- Eric Bogert, Aaron Schecter, and Richard T Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. Scientific reports, 11(1):8028, 2021.
- Svetlin Bostandjiev, John O’Donovan, and Tobias Höllerer. Tasteweights: a visual interactive hybrid recommender system. In Proceedings of the sixth ACM conference on Recommender systems, pages 35–42, 2012.
- Paul A Bottomley and John R Doyle. Comparing the validity of numerical judgements elicited by direct rating and point allocation: Insights from objectively verifiable perceptual tasks. European Journal of Operational Research, 228(1):148–157, 2013.
- Paul A Bottomley, John R Doyle, and Rodney H Green. Testing the reliability of weight elicitation methods: direct rating versus point allocation. Journal of Marketing Research, 37(4):508–513, 2000.
- Stijn Broecke. Oecd social, employment and migration working papers: Artificial intelligence and labour market matching. 2023.
- Michèle Bélot, Philipp Kircher, and Paul Muller. How wage announcements affect job search: A field experiment. IZA Discussion Papers, 11814, September 2018.
- Michèle Bélot, Philipp Kircher, and Paul Muller. Providing advice to jobseekers at low cost: An experimental study on online advice. The Review of Economic Studies, 86(4): 1411–1447, 2019.

- Michèle Bélot, Philipp Kircher, and Paul Muller. Do the long-term unemployed benefit from automated occupational advice during online job search? SSRN Electronic Journal, 01 2022. doi: 10.2139/ssrn.4178928.
- Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. Journal of Marketing Research, 56(5):809–825, 2019.
- Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Fisherschlutz lecture: Generic machine learning inference on heterogenous treatment effects in randomized experiments, with an application to immunization in india, 2023. URL <https://arxiv.org/abs/1712.04802>.
- Hyesun Choung, John S Seberger, and Prabu David. When ai is perceived to be fairer than a human: Understanding perceptions of algorithmic decisions in a job application context. International Journal of Human–Computer Interaction, pages 1–18, 2023.
- Michael Cooper and Peter Kuhn. Behavioral job search. Handbook of Labor, Human Resources and Population Economics, pages 1–22, 2020.
- Peter A Diamond and Jerry A Hausman. Contingent valuation: is some number better than no number? Journal of economic perspectives, 8(4):45–64, 1994.
- Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1):114, 2015.
- Sara Drenner, Shilad Sen, and Loren Terveen. Crafting the initial user experience to achieve community goals. In Proceedings of the 2008 ACM conference on Recommender systems, pages 187–194, 2008.
- European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, April 2021. COM(2021) 206 final.
- Brian Feld, AbdelRahman Nagy, and Adam Osman. What do jobseekers want? comparing methods to estimate reservation wages and the value of job attributes. Journal of Development Economics, 159:102978, 2022a.
- Brian Feld, AbdelRahman Nagy, and Adam Osman. What do jobseekers want? comparing methods to estimate reservation wages and the value of job attributes. Journal of

- Development Economics, 159:102978, 2022b. ISSN 0304-3878. doi: <https://doi.org/10.1016/j.jdeveco.2022.102978>. URL <https://www.sciencedirect.com/science/article/pii/S0304387822001201>.
- Erica Field, Robert Garlick, Nivedhitha Subramanian, and Kate Vyborny. Why don't job-seekers search more? barriers and returns to search on a job matching platform. Working paper, 2023.
- Mauricio N Freire and Leandro N de Castro. e-recruitment recommender systems: a systematic review. Knowledge and Information Systems, 63:1–20, 2021.
- Francisco Gutiérrez, Sven Charleer, Robin De Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems. In Proceedings of the 13th ACM Conference on Recommender Systems, pages 60–68, 2019.
- Gerald Häubl and Valerie Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. Marketing science, 19(1):4–21, 2000.
- Roland Helm, Armin Scholl, Laura Manthey, and Michael Steiner. Measuring customer preferences in new product development: comparing compositional and decompositional methods. International Journal of Product Development, 1(1):12–29, 2004.
- Gunter J Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. American Economic Review, 100(1):130–63, 2010.
- John J Horton. The effects of algorithmic labor market recommendations: Evidence from a field experiment. Journal of Labor Economics, 35(2):345–385, 2017.
- Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. 2020.
- Maximilian Kasy. The political economy of ai: Towards democratic control of the means of prediction. SSRN Electronic Journal, 2024.
- Erin M Kelley, Christopher Ksoll, and Jeremy Magruder. How do digital platforms affect employment and job search? evidence from india. Journal of Development Economics, 166:103176, 2024.
- Philipp Kircher. Job search in the 21st century. Journal of the European Economic Association, 20(6):2317–2352, 2022.

- Philipp AT Kircher. Search design and online job search—new avenues for applied and experimental research. Labour economics, 64:101820, 2020.
- Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences, 118(22), 2021.
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. arXiv preprint arXiv:2202.11776, 2022.
- Jon A Krosnick and Duane F Alwin. A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. Public Opinion Quarterly, 52(4): 526–538, 1988.
- Peter Kuhn and Hani Mansour. Is internet job search still ineffective? The Economic Journal, 124(Issue 581):1213–1233, December 2014a. doi: 10.1111/eoj.12119. URL <https://onlinelibrary.wiley.com/doi/full/10.1111/eoj.12119>.
- Peter Kuhn and Hani Mansour. Is internet job search still ineffective? The Economic Journal, 124(581):1213–1233, 2014b.
- Sven Laumer, Fabian Gubler, Christian Maier, and Tim Weitzel. Job seekers’ acceptance of job recommender systems: Results of an empirical study. 2018.
- Thomas Le Barbanchon, Roland Rathelot, and Alexandra Roulet. Gender differences in job search: Trading off commute against wage. The Quarterly Journal of Economics, 136(1):381–426, 2021.
- Thomas Le Barbanchon, Lena Hensvik, and Roland Rathelot. How can ai improve search and matching? evidence from 59 million personalized job recommendations. Technical report, Working Paper, 2023.
- Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1): 2053951718756684, 2018.
- Danielle Li, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Technical report, National Bureau of Economic Research, 2020.
- Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. J. Mach. Learn. Res., 20(80):1–37, 2019.

- Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90–103, 2019.
- Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4):629–650, 2019.
- Nicole Maestas, Kathleen J Mullen, David Powell, Till Von Wachter, and Jeffrey B Wenger. The value of working conditions in the united states and implications for the structure of wages. American Economic Review, 113(7):2007–2047, 2023.
- Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175:121390, 2022.
- Charles F Manski, Kenneth I Wolpin, and Elke U Weber. Analysis of choice expectations in incomplete scenarios. Elicitation of Preferences, pages 49–72, 2000.
- Kevin Marsh, Maarten IJzerman, Praveen Thokala, Rob Baltussen, Meindert Boysen, Zoltán Kaló, Thomas Lönngren, Filip Mussen, Stuart Peacock, John Watkins, and Nancy Devlin. Multiple criteria decision analysis for health care decision making—emerging good practices: Report 2 of the ispor mcda emerging good practices task force. Value in Health, 19(2):125–137, 2016. ISSN 1098-3015. doi: <https://doi.org/10.1016/j.jval.2015.12.016>. URL <https://www.sciencedirect.com/science/article/pii/S1098301515300152>.
- Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. American Economic Review, 107(12):3722–59, 2017a.
- Alexandre Mas and Amanda Pallais. Valuating alternative work arrangements. American Economic Review, 107(12):3722–3759, 2017b.
- Yoosof Mashayekhi, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. Recon: Reducing congestion in job recommendation using optimal transport. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 696–701, 2023.
- Yoosof Mashayekhi, Nan Li, Bo Kang, Jefrey Lijffijt, and Tijl De Bie. A challenge-based survey of e-recruitment recommendation systems. ACM Computing Surveys, 56(10): 1–33, 2024.
- John A McCarty and Larry J Shrum. The measurement of personal values in survey research: A test of alternative rating procedures. Public Opinion Quarterly, 64(3):271–298, 2000.

- Andreas I Mueller and Johannes Spinnewijn. Expectations data, labor market, and job search. Handbook of Economic Expectations, pages 677–713, 2023.
- Andreas I Mueller, Johannes Spinnewijn, and Giorgio Topa. Job seekers’ perceptions and employment prospects: Heterogeneity, duration dependence, and bias. American Economic Review, 111(1):324–363, 2021.
- Jessica Ochmann, Leonard Michels, Sandra Zilker, Verena Tiefenbeck, and Sven Laumer. The influence of algorithm aversion and anthropomorphic agent design on the acceptance of ai-based job recommendations. In ICIS, 2020.
- Denis Parra and Peter Brusilovsky. User-controllable personalization: A case study with setfusion. International Journal of Human-Computer Studies, 78:43–67, 2015.
- Florian Pethig and Julia Kroenung. Biased humans, (un) biased algorithms? Journal of Business Ethics, 183(3):637–652, 2023.
- Andrew Prael and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting, 36(6):691–702, 2017.
- Pearl Pu, Li Chen, and Rong Hu. Evaluating recommender systems from the user’s perspective: survey of the state of the art. User Modeling and User-Adapted Interaction, 22:317–355, 2012.
- Kirsten Swearingen and Rashmi Sinha. Beyond algorithms: An hci perspective on recommender systems. In ACM SIGIR 2001 workshop on recommender systems, volume 13, pages 1–11, 2001.
- Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. My friends, editors, algorithms, and I: Examining audience attitudes to news selection. Digital journalism, 7(4):447–469, 2019.
- Bas van der Klaauw and Heike Vethaak. Empirical evaluation of broader job search requirements for unemployed workers. Technical report, Tinbergen Institute Discussion Paper, 2022.
- Patrick van Esch, J Stewart Black, and Denni Arli. Job candidates’ reactions to ai-enabled job application processes. AI and Ethics, 1:119–130, 2021.
- Koert Van Ittersum, Joost ME Pennings, Brian Wansink, and Hans CM Van Trijp. The validity of attribute-importance measurement: A review. Journal of Business Research, 60(11):1177–1190, 2007.

- Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. MIS quarterly, pages 425–478, 2003.
- Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 - RecSys Challenge 17. ACM Press, 2017. doi: 10.1145/3124791.3124792.
- Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In Proceedings of the 2020 CHI conference on human factors in computing systems, pages 1–14, 2020.
- Marieke GM Weernink, Sarah IM Janus, Janine A Van Til, Dennis W Raisch, Jeannette G Van Manen, and Maarten J IJzerman. A systematic review to identify the use of preference elicitation methods in healthcare decision making. Pharmaceutical medicine, 28: 175–185, 2014.
- Jenny S Wesche and Andreas Sonderegger. Repelled at first sight? expectations and intentions of job-seekers reading about ai selection in job advertisements. Computers in human behavior, 125:106931, 2021.
- Matthew Wiswall and Basit Zafar. Preference for the workplace, investment in human capital, and gender. The Quarterly Journal of Economics, 133(1):457–507, 2017.
- Matthew Wiswall and Basit Zafar. Preference for the workplace, investment in human capital, and gender. The Quarterly Journal of Economics, 133(1):457–507, 2018.
- World Bank. The use of advanced technology in job matching platforms: Recent examples from public agencies. 2023. URL <https://thedocs.worldbank.org/en/doc/ceb5c5792ad0d874e9b1c3cc71362f46-0460012023/original/Digital-Job-Matching-Platforms-S4YE-Draft-Note-for-Discussion.pdf>.
- Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. Journal of Behavioral Decision Making, 32(4):403–414, 2019.
- Noorul Hassan Zardari, Kamal Ahmed, Sharif Moniruzzaman Shirazi, and Zulkifli Bin Yusop. Weighting methods and their effects on multi-criteria decision making model outcomes in water resources management. Springer, 2014.

Jing Zhao, Jingya Wang, Madhav Sigdel, Bopeng Zhang, Phuong Hoang, Mengshu Liu, and Mohammed Korayem. Embedding-based recommender system for job to candidate matching on scale. arXiv preprint arXiv:2107.00221, 2021.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of Misaligned AI. Advances in Neural Information Processing Systems, 33:15763–15773, 2020.

Chapter 1

Designing Job Recommender Systems: How to Improve Human-Based Search

This chapter is a joint work with Guillaume Bied (University of Ghent), Philippe Caillou (UP-Saclay - LISN - INRIA), Bruno Crépon (CREST), Christophe Gaillac (University of Geneva), and Michèle Sebag (UPSaclay - LISN - INRIA - CNRS).¹

Abstract: There are various possible Recommendation Systems (RS), typically optimized using algorithmic criteria with limited consideration of individual job search behavior. This article explores how economic theory can help identify an optimal RS from the perspective of job seekers. We demonstrate how a job search model can be used to design such RS. The model is empirically validated using a large set of administrative job search data. We then deduce the form of the optimal RS and compare its potential performance to that of conventional RS models.

¹This research project is the result of a partnership with France Travail, the Public Employment Service in France. We thank Paul Beurnier, H el ene Caillol, Pierre-Antoine Corre, Yann De Coster, Thierry Foltier, Cyril Nouveau, Camille Qu er e, S ebastien Robidou, and Chantal Vessereau for their operational support. This research was supported by DATAIA convergence institute as part of the “Programme d’Investissement d’Avenir”, (ANR-17-CONV-0003) operated by CREST and LISN. The authors retained full intellectual freedom throughout this process, and any errors are our own.

Introduction

In recent years, there has been a surge in the use of recommendation algorithms in the labor market, reflecting the belief that the internet can enhance labor market efficiency and job matches' quality [Autor, 2001, Kuhn and Mansour, 2014, Horton, 2017]. A variety of tools and approaches to generate such recommendations have been developed. Platforms and Public Employment Services (PESs), for example, have implemented tools to recommend job vacancies aligning with job seekers' preferences (see Broecke, 2023, Gutiérrez et al., 2019). Advanced Machine Learning (ML) tools have also emerged, proposing different types of recommendation systems, typically optimized using algorithmic criteria with limited consideration of individual behavior (RS hereafter, see, e.g., Freire and de Castro [2021] and proceedings of RecSys, which is a cycle of ML conferences exclusively focusing on issues related to RS). Indeed, while ML algorithms are adapted to different observational contexts, they primarily aim to predict the success of a match in various sense (predicting the hiring likelihood, of applying for a job, or even job seekers' interest).¹ The underlying mechanism behind the potential efficiency gains of these algorithms is that they perform faster and at lower cost screening tasks that job seekers would do themselves.

At the same time, a number of studies have sought to measure and investigate the added value that RSs can bring. The existence of job search platforms and the possibility of large-scale experiments have also become a powerful means of studying search behavior and ways of improving it [Kircher, 2020, 2022]. While some studies are devoted to the recommendation of job seekers to recruiters (see for example Horton [2017] and Li et al. [2019]), many studies focus on recommendations to job seekers. Le Barbanchon et al. [2023] and Field et al. [2023] focus on recommendations of job vacancies to job seekers. Behaghel et al. [2024] considers firm recommendations to job seekers. An important area that has been studied is the recommendation of job seekers to broaden their search and consider other labor markets (Belot et al. [2019, 2022], Altmann et al. [2022]). In this case, the underlying mechanism is that RS can redirect search efforts towards types of job vacancies where they would be more productive.

Several important questions emerge from the growing use of recommendation systems, particularly from the perspective of job seekers. A key issue is understanding the

¹Some algorithms, known as collaborative filtering strategies, rely solely on past choices to recommend jobs that other job seekers, who previously applied for the same jobs as the current user, have chosen (see, e.g., Le Barbanchon et al., 2023). Other algorithms, so-called *content-based recommenders* or *hybrid*, leverage comprehensive data about jobs, job seekers, and past choices [see, e.g., Volkovs et al., 2017, Zhao et al., 2021, Bied et al., 2021b, Mashayekhi et al., 2023]. They aim to predict matches by considering job seeker and vacancy characteristics. Some algorithms, instead, construct similarity measures based solely on the attributes of job seekers and job vacancies, without considering historical choices or matches.

diversity of RS and identifying the conditions under which an RS effectively enhances a job seeker’s outcomes. Central to this is the role of economic models in capturing the complexities of job search behavior. Another closely related challenge is determining whether there exists an optimal RS tailored to job seekers’ needs, and if so, how such a system should be designed. This requires not only observing the choices made by job seekers and firms but also accurately modeling and estimating their underlying behavior. Finally, it is critical to assess whether simply replicating job seekers’ past choices is a sufficient foundation for RS design, even when those choices are rational [see more broadly, e.g., Zhuang and Hadfield-Menell, 2020, Kleinberg and Raghavan, 2021, Kleinberg et al., 2022, Kasy, 2023, on value misalignment problem].

Other important questions also arise. First, if RSs improve job prospects by narrowing the set of relevant opportunities, they are likely to make job seekers more selective, as demonstrated by Kelley et al. [2024]. How does this increased selectivity influence the overall benefits derived from RSs? Should this behavioral change be factored into the design of these systems? Furthermore, it is essential to consider potential biases in job seekers’ perceptions of their likelihood of success. As discussed by Belot et al. [2019, 2022], Altmann et al. [2022], using RSs to redirect job search efforts to areas where they would be more productive is a promising strategy. However, how do these biases—along with the need to recommend vacancies that job seekers find appealing—constrain the design of effective RSs that truly benefit both job seekers and employers?

Leveraging large-scale administrative data that offer a uniquely comprehensive view of the search and matching process—encompassing the characteristics of job vacancies, job seekers’ search parameters across various job dimensions, and their multiple interactions (e.g., clicks, applications, and hires)—this paper aims to address these questions. We begin by highlighting the empirical significance of these issues, showing that different RSs capture distinct dimensions of the job search process and generate notably different recommendations. Specifically, we analyze two algorithms. The first, commonly used by PESs, matches job vacancies to job seekers based on declared search preferences. The second, reflective of a growing body of machine learning literature, predicts successful matches by analyzing job seeker and vacancy characteristics. The rationale behind these designs, which we validate empirically, is that the first algorithm captures the utility \mathcal{U} a job seeker derives from a vacancy, while the second captures the recruitment chances \mathcal{P} of an application. We demonstrate that these two approaches yield highly divergent sets of recommendations in practice. To assess the potential of combining these approaches, we conducted a field test within the French PES. Job seekers were randomly assigned to receive recommendations from either the preference-based or ML-based algorithms, as well as from hybrid models that integrate both methods. The results show that while ML-

based recommendations enhance perceived hiring chances, hybrid models that balance both hiring probabilities and job preferences generate higher engagement, particularly through increased click-through rates. These findings suggest that integrating both dimensions improves the relevance and effectiveness of job recommendations.

We develop a model to define the value of search when using a RS. The model incorporates two key factors: the utility of a job (U) and the chances of success of an application (P). This model helps address several of the previously raised questions. First, we show that there exists an optimal RS, and that optimal recommendations are based on a score that combines both U and P . Interestingly, while there may be ML-based algorithms to assess P (such as the one we employ, denoted by \mathcal{P}), no such algorithm exists to assess U . In fact, the version we use, denoted as \mathcal{U} , is derived from an expert system rather than an ML-based score. Second, combining these two factors, we demonstrate that the functional form of the optimal RS score cannot be learned solely by observing and replicating job seekers' behavior. The model also outlines the conditions under which an RS improves a job seeker's search value. Furthermore, the model shows that application behavior depends on the two previously mentioned factors, U and P , in a manner that closely aligns with the findings of [Hitsch et al. \[2010\]](#) in the context of online dating. By interpreting \mathcal{U} and \mathcal{P} as representations of these two key dimensions, we estimate the application model using job seekers' application data. Consistent with the model, we find that these two variables significantly influence the decision to apply for posted jobs. This constitutes the first empirical result of the paper, validating the model and the two score representations, and providing evidence that both are central to job seekers' decision-making processes.

The estimates derived above allow us to reconstruct the valuation functions for specific vacancies for each job seeker, as well as to determine the optimal recommendation system (RS). Using these results, we evaluate the recommendation sets produced by the two initial RSs—based on the \mathcal{U} and \mathcal{P} scores—and the optimal algorithm. By comparing these values to a benchmark that reflects job seekers' own searches, while holding screening effort constant, we are able to quantify the value added by each RS. This forms the second key empirical contribution of the paper.

Our findings reveal that the optimal RS and the ML-based RS (driven by \mathcal{P}) perform similarly, both adding substantial value to the job search process for job seekers. In contrast, the algorithm relying on matching search parameters (i.e., \mathcal{U}) performs significantly worse. In fact, it leads to a decrease in the value of job seekers' search efforts when compared to the benchmark.

Importantly, we find that the performance of the optimal RS varies widely across individuals. While the optimal RS substantially improves outcomes for job seekers with strong labor market prospects, the benefits are far more modest for those with weaker

prospects. This heterogeneity underscores the fact that RSs, while adding value overall, may exacerbate existing labor market inequalities by disproportionately benefiting those already well-positioned.

Overall, our study underlines the importance of identifying several essential quantities for building a high performance RS, in particular the job utility (U) and the probability of a successful application (P). The latter can be derived from the available data, a task at which ML tools excel, while exploiting the complexity of the information available on both sides of the market. However, identifying a prediction of the job seekers' utility associated with a position, seemingly simpler, is challenging due to the lack of direct observation.

Related literature

In essence, this paper relates different strands of the literature. Given the importance of online job search [Kuhn and Mansour, 2014, Kircher, 2022], the first is related to the impact of recommendations on labor market frictions. Following Belot et al. [2019], several studies show that suggestions to expand the search to alternative occupations have an effect on interviews and future job outcomes [Belot et al., 2022, Altmann et al., 2022]. Relatedly, some studies discuss how the identification of sectors to be recommended can be refined by collecting specific data on skills from job seekers [Bächli et al., 2024]. While in the previous papers, broadening job search to recommended job markets is a suggestion, sometimes taking the time of directly showing vacancies as in Belot et al. [2019], broadening job search can be made compulsory. van der Klaauw and Vethaak [2022] show that in such a case, the impact can be negative on both job search and job quality, which might suggest heterogeneous impact and room for improving recommendations.² Other studies instead recommend firms predicted to hire in a broaden labor market: Behaghel et al. [2024] show in a randomized experiment in which job seekers are randomly assigned to treatment as well as the firms recommended to them, that being assigned to receive recommendations has an impact, although small, on hiring, mainly at recommended firms. Drawing on the specifics of their experimental design, they show that recommended firms indeed and as expected have hiring potential conditional on application, and that the treatment's effect on job seeker-firm matching can be attributed to an increase in the application rate on recommended pairs. Other studies instead use job seekers' interest in certain vacancies, using click history to generate recommendations: Le Barbanchon et al. [2023] shows in a two-sided experiment that such RS redirects job seekers' search and hiring towards recommended vacancies, which themselves receive more applications. As these studies show, and sometimes discuss [Behaghel et al., 2024],

²In this study, broadening the job search means applying for jobs in other occupations or geographical areas, offering a lower salary or requiring lower skills.

there are several conditions for these RS to be effective. They need to arouse the interest of job seekers, but also direct them towards vacancies with good recruitment potential. And as they show, it is not out of the question that recommendations may lead to a deterioration in the situation of job seekers. It is worth to note that these studies use very different strategies to generate recommendations. Our contribution is to illustrate firstly that recommendations based on different principles do, in fact and quite logically, generate highly different recommendations, both in terms of chances of recruitment and their attractiveness from the job seeker's point of view. Our contribution is then, thanks to a data-validated theoretical model, to identify the conditions necessary for a RS to improve the situation of job seekers, and to define one RS that is optimal from their point of view. One key insight of our analysis is that RS often focus on a precise objective - typically improving the chances of a match - which may differ from the job seekers' objectives. This disconnection between the two can result in substantial losses, as some individuals may focus their search on vacancies far from their preferences following the recommendations.

In addition, the literature highlights the importance of taking into account the behavioral aspects of the job search [see, e.g., Babcock et al., 2012, Cooper and Kuhn, 2020, Altmann et al., 2018], whether they be, among others, biased perceptions of the chances of success of search strategies or biased perceptions of market fundamentals. Indeed, recent empirical work has highlighted the existence of a category of job seekers who remain permanently over-optimistic about their chances of returning to work [Mueller et al., 2021, Mueller and Spinnewijn, 2023]. These biases could prevent job seekers from applying for high-yield vacancies: Field et al. [2023] show in an experiment involving recommendations made to job seekers, that the inertia of behavior due to the simple fact of calling a firm to apply for a potentially interesting recommended vacancy can be a major barrier to applications. They also show that lowering this psychological cost greatly increases the number of applications. More directly, as shown by Kelley et al. [2024], being exposed to a better pool of vacancies may lead job seekers, as predicted by the basic search model, to adjust their search behavior by being more selective about which vacancies to apply for. Our contribution is, within the framework of a sufficiently flexible model, to discuss how behavioral adjustment or the existence of behavioral biases can affect the effectiveness of RS and how they could be taken into account in its design. Our paper highlights the importance of designing RS based on both individual hiring prediction and a relevant and detailed representation of behaviors, starting for example, with the importance of accurately measuring preferences for different job attributes [Mas and Pallais, 2017, Wiswall and Zafar, 2018, ?, Banerjee et al., 2022, Banerjee and Chiplunkar, 2024].

Focusing on this central aspect of basing RS on job seekers' behaviors, our paper leaves in the shade the important dimension of labor market congestion and the impact of RS

on congestion. Yet congestion is a key dimension of the functioning of RS. It is central to several of the papers cited above Kircher [2022], Behaghel et al. [2024], Le Barbanchon et al. [2023] which, despite large-scale randomized experiments, sometimes fail to provide evidence on such congestion effects. Congestion is also instrumental in our previous work Bied et al. [2021b,a], which sketches the path of optimal transport as a means of dealing with this problem. In this contribution, we focus on the instrumental dimension of behavior to design effective RS, with the feeling that potential issues that congestion may pose when designing RS will be more easily and effectively studied once an effective recommender score has been developed.

The remainder of the paper is structured as follows. In Section 1, we provide a comprehensive overview of the job RSs under consideration, including the machine learning (ML)-based RS and the preference-based RS. This section also introduces the dataset used in the analysis. Section 2 offers a detailed analysis of the two representative RSs implemented in our study and examine their respective recommendation policies, emphasizing their distinct approaches to matching job seekers with vacancies. Section 3 presents the theoretical framework, which models the job search behavior when using an RS. Finally, in Section 4, we empirically evaluate the performance of the two RSs using the evaluation metrics from the theoretical model, and compare them with the benchmark search process where job seekers manually select vacancies.

1 Study context

1.1 Review of existing job RSs

All recommendation systems work in a similar way. They are based on the computation of a score: for an individual i characterized by variables x_i and a vacancy j characterized by variables y_j , there is a score $S(i, j)$ depending on x_i and y_j such that higher score values are preferred. The derivation of a set of recommendations from a matching score $S_{i,j}$ is straightforward. For a job seeker i_0 , the score is used to rank existing vacancies from the most to the least desirable. To make k recommendations to i_0 , an intuitive solution is to pick the k vacancies with the largest score $S_{i_0,j}$.

Although the principle is the same, there is a variety of approaches to job recommendation which exist in the computer science literature and in applications, as surveyed by Freire and de Castro [2021], De Ruijt and Bhulai [2021], Mashayekhi et al. [2022]. This reflects a multitude of application contexts, available datasets and algorithmic strategies, as summarized in Table 1.1.

Knowledge-based RSs leverage expert knowledge of the labor market, captured in detailed ontologies (e.g. of jobs, skills, contracts) and in relationships between their entities, to match people and jobs based on assessed fit quality. A prominent example is the WCC ELISE matching solution, used by several national PESs as well as private entities such as Robert Half.³

A variety of approaches instead leverage machine learning techniques to recommend jobs. On the one hand, *collaborative filtering* strategies exclusively rely on individual interaction histories of job seekers with job ads, in order to define similarities that can be leveraged to show job seekers job ads that are similar (in terms of browsing patterns) to the ones they have clicked on in the past. An example is the matrix factorization algorithm studied by [Le Barbanchon et al. \[2023\]](#) in the context of the Swedish PES. On the other hand, what could be called *content-based* RSs primarily leverage descriptions of the characteristics of job seekers and job ads (e.g. in terms of jobs, education and skills) to generate recommendations. Such systems may for instance predict the probability of a specific type of job seeker-job ad interaction (clicks, applications, hires) given job seekers' and job ads' descriptions x_i and y_i , in order to show job seekers the most promising job ads in terms of estimated interaction probability. An example is CareerBuilder's job RS [[Zhao et al., 2021](#)].

Hybrid RS combine several of the aforementioned approaches. The winning approach in the RecSys 2017 job recommendation challenge [[Volkovs et al., 2017](#)] predicts interactions by utilizing user and item features, along with hand-crafted features that compare job ads to those already viewed by job seekers. LinkedIn's RS predicts matches based on user and job ad characteristics, enhancing personalization with individual and recruiter-level fixed effects when sufficient interactions are present [[Shi et al., 2022](#)]. Indeed's RS involves collaborative filtering and content based systems, post-processed by a hybrid approach involving content-based deep learning and a rule-based engine [[Ma et al., 2022](#)].

Although algorithms can be obtained according to different principles and with different data, there is a common way of measuring their performance, which is the "Recall@k". Let's consider a target variable M , such as $M_{i,j} = 1$ if there i has been hired by j . As usual the algorithm if trained, is trained on a "train sample", and tested on a "test sample". For each individual i and the k best vacancies according to S vacancy in the test sample $\mathcal{J}_k^*(S, i)$, we can build a variable $M_i^k(S)$ which takes value 1 if the target variable $M_{i,j}$ takes value 1 for one of these k best S -based vacancies:

³See the dedicated sites [Robert Half](#) and [ELISE](#).

Table 1.1: Some examples of different RSs

References	Setting	Knowledge based	Collaborative Filtering	Content based	Target variable
WCC Elise	National PESs, Robert Half	x			
Le Barbanchon et al. [2023]	Swedish PES		x		Clicks
Zhao et al. [2021]	CareerBuilder			x	Applications
Shi et al. [2022]	LinkedIn		x	x	Applications, "save"
Volkovs et al. [2017]	Xing challenge		x	x	Impressions, clicks
Ma et al. [2022]	Indeed	x	x	x	Clicks, Applications

$$M_i^k(S) = \mathbb{1} \left(\sum_{j \in \mathcal{J}_k^*(S,i)} M_{i,j} = 1 \right), \quad (1.1)$$

where $\mathbb{1}(\cdot)$ denotes the the indicator function. If the target variable is hiring, the $\text{recall@}k$ is the proportion of job seekers who were hired on one of the top- k recommendations:

$$\text{recall@}k(S) = \frac{1}{N} \sum_{i=1}^N M_i^k(S). \quad (1.2)$$

This is the usual measure in the machine learning literature of the global performance of the RS S , which can be used for example to compare two RSs.

1.2 A machine learning RS based on hiring predictions

We now describe the job RS that we developed [see Bied et al., 2021b, 2023b, for more details on the architecture and the related literature] which is a state-of-the-art RS building on the insights of the winning algorithm of the RecSys 2017 challenge [Volkovs et al., 2017]. An originality of our recommendation algorithm lies in the embedding related to the geographical distance of the job seeker and job ad that we describe below. To evaluate the models and select their hyperparameters selected, an objective would be to optimize directly the so-called $\text{recall@}k$ on the test set: the proportion of job seekers i in the test sample who where hired on a vacancy j in the top k recommendations (among vacancies available the week of the match), where k is usually 10, 20, 50, or 100. However, while the $\text{recall@}k$ defines a performance metric to evaluate algorithms on the test set, it is intractable for direct optimization. Drawing on the learning to rank literature, we propose to learn a similarity score $S_{i,j}$ between job seeker i and job ad j as a function of the job seeker's and the ad's characteristics. Given a job seeker i and two vacancies j and j' , we wish S_{ij} to be lower than $S_{ij'}$ if i matched with j and not with j' . This motivates the min-

imization on the training set of the so-called *triplet margin loss* [see, e.g., Weinberger and Saul, 2009] corresponding to the following objective:

$$\min_S \sum_i \sum_{j' \neq j^*(i)} [S_{i,j^*(i)} - S_{i,j'} + \eta]_+, \quad (1.3)$$

where $\eta > 0$ is a scalar hyperparameter, $[x]_+ = \max(x, 0)$, the outer sum ranges over all job seekers with matches, the inner one over all job ads, and $j^*(i)$ is the job ad with which i actually matched. This expression aims at separating, for all job seekers, the scores associated to ads they matched with from the scores associated to all other job ads by a margin of at least η .

Given job seeker and job ad characteristics, resp. X_i and Y_j , the score S_{ij} is defined as:

$$S_{i,j}(X_{i,j}) = \phi(X_i)^T A \psi(Y_j),$$

where $X_{i,j} = (X_i, Y_j)$, ϕ, ψ are feed-forward neural networks with several layers, and A is an affinity matrix. Feed-forward neural networks are flexible, differentiable functions commonly used in the machine learning literature, that handle high-dimensional features well when given large datasets.⁴

In this context, $\phi(X_i)$ and $\psi(Y_j)$ may be understood as latent variables describing i and j . A can be interpreted as an affinity matrix: the parameter $A_{k,l}$ represents the complementarity between dimension k of the job seekers' latent space and dimension l of the ad's latent variables. The latent space is of size 872 for both job seekers and vacancies, although ϕ, ψ and A are given a block-wise structure which incorporates domain knowledge (three blocks corresponding to geography, to skills, to other factors) and reduces the number of parameters. In other words, and as schematically represented on Figure 1.1,

$$S_{i,j}(X_{i,j}) = \sum_{k \in \{ \text{"geography"}, \text{"skills"}, \text{"other features"} \}} \phi_k(X_i)^T A_k \psi_g(Y_j).$$

The parameters that are optimized during training are the parameters of the transformations from the observed to the latent variables (*i.e.* the weights of the neural networks) and the affinity matrices. As is customary in the machine learning literature, the minimization of this (non-convex) objective function is done by mini-batch stochastic gradient descent. For computational efficiency, pairs that are not matches are very aggressively uniformly subsampled.

⁴The interested reader may consult Goodfellow et al. [2016] for a textbook treatment.

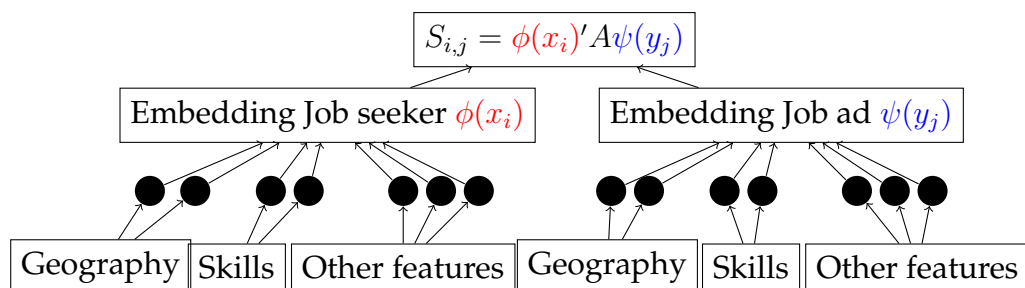


Figure 1.1: High-level Structure of the neural network

used to describe the similarity between users and items, using a bilinear form between their respective embeddings.

1.3 A preference-based RS based on search criteria

The French PES has developed a matching algorithm based on WCC Elise, which is used to suggest relevant vacancies to job seekers. The principle behind such RSs is to take the characteristics of the job desired by job seekers and the jobs available as a starting point. For each characteristic taken into account, a sub-score is determined, ranging from zero to one, depending on the degree of match between the characteristics of the job available and those of the desired job. The sub-scores are then aggregated to form a global score. This global score is obtained mainly by considering a weighted average of the sub-scores, but also by imposing that the nullity of certain sub-scores extends to the global score.

For the purpose of this study we built a RS inspired by the one used at the PES and based on the same list of criteria. But as we only had partial access to the RS used by the PES, it is likely to be different. In practice, the selected criteria are matched with the same exact criteria on the recruiters' side (i.e., profile required in the vacancy and characteristics of the proposed job). For each characteristic k , a consistency criterion $c_k(i, j) \in [0, 1]$ is defined, corresponding to whether characteristic k of the job seeker profile is consistent with the characteristic in the vacancy. For example, for "reservation wage", the criterion takes the value 1 if the wage offered in the vacancy is above the reservation wage in the job seeker's profile. For "geographic mobility", the criterion takes the value 1 if the distance between the job location and the job seeker home-place is below the job seeker's maximum commuting distance. The exact criteria used at the PES share the same principles but allow for more smoothness in the definition of the sub-criteria used for each characteristic.

"Skills" and "Occupation" are important criteria. In his/her individual profile, a job seeker enters a primary occupation she is searching for, as well as a set of skills. The

criterion “Occupation” takes value 1 if the occupation the job seeker seeks is the exact same occupation as the one entered in the vacancy, or if it is a close occupation. This proximity is defined according to an expert-based matrix made available by the PES. The criterion “skills” corresponds to the adequacy between the skills listed in an individual profile and the skills listed in the vacancy.

These criteria can be seen as components of the utility function of job seekers. Criteria like “Skills”, “Diploma”, “Year of experience”, “Driving Licence”, “Languages” can enter the utility as they relate more precisely than the occupation to the precise skill set that a job seeker would be able to use on the advertised job. Indeed, the fit with ones’ skill set is related to personal fulfillment and development of human capital on-the-job, which is valued by job seekers.

Each criterion is then associated with a weight w_k and the final matching score is the weighted sum of each single fit between the criteria of applicants and the job ad’s content. The score used at the PES involves some nonlinearities that we ignore to maintain the interpretation. The simplified version we use is:

$$\mathcal{U}(i, j) = \sum_{k=1}^K w_k c_k(i, j), \quad (1.4)$$

where the set of normalized weights $\{w_k\}_{k=1, \dots, K}$ is the same for the whole population, namely Occupation (0.332), Skills in occupation (0.332), Geographic mobility (0.997), Reservation wage (0.066), Diploma (0.033), Working hours (0.033), Driving license (0.033), Languages (0.033), Years of experience in occupation (0.033), Duration and type of contract (0.003).

1.4 Data

We have access to rich historical administrative data from the PES from weeks 1 to 48 of 2019, which allows us to train and evaluate several job recommendation systems. This includes descriptions of vacancies posted on the PES, characteristics and search parameters of job seekers, as well as clicks on the PES website, applications, and hiring data. We restrict ourselves to the former French region of Rhône-Alpes, which is sufficiently diverse economically and geographically. In this market, we therefore use data on 1,181,902 (or 516,776) unique job seeker search sessions (or job ads); and on average, 610,986 job seekers (or 129,642 job ads) are active in a given week.

The PES website is open to all employers and job seekers and is the largest outlet

for vacancies on the French labour market.⁵ The data include many characteristics of these job postings such as the date of the publication, the occupation at several levels of granularity, the salary offered, the experience required, the type of contract (permanent, temporary or fixed-term), the location of the workplace, the weekly working hours, the qualifications required, but also the hard and soft skills (as reduced by singular value decomposition as well as raw data), the textual descriptions of the job posting and of the firm (reduced by singular value decomposition), the size of the firm, the number of applications to the advertisement and to the firm in the last six months, the time since the vacancy was published, etc.

Demographic and search information on the unemployment spells of job seekers come from the administrative files (*e.g.*, the *fichier historique* (FH) of the PES). Importantly, they include date of registration, geographic location, experience, skills, duration of unemployment, applications in the last six months, various individual and postal code level socio-demographic characteristics. We also know about job search parameters, such as those declared when registering with the PES: reservation wage, maximum commuting time, desired job, desired type of contract (temporary vs. long-term), and working hours (full-time vs. part-time).^{6,7}

This comprehensive information on both sides of the market is complemented by the clicks on the PES website and the applications made by these job seekers on these job postings. If this application data has been used in [Marinescu and Skandalis \[2021\]](#), [Glover \[2019\]](#), [Algan et al. \[2020\]](#), the use of clicks to measure the interest in job vacancies in the preliminary phase of the search on the website seems to be new. The application data consist of applications through three of the PES channels: applications made directly by job seekers, potential matches initiated by the firm, and suggestions initiated by the PES case workers. We observe 75,744 successful matches in the data. Finally, we also exploit the final outcome of these interactions, whether or not there is a hire. This information is recorded by the caseworkers.⁸

Observations from week 1 to 43 of 2019 are used as a training set (representing 66,914 matches) for the two RSs; while weeks 44 to 48 (representing 8,830 matches) are used as a test set to evaluate the quality of recommendations.

⁵According to [Le Barbanchon et al. \[2021\]](#), which uses the same data on applications, they represented 60% in 2010, see Section VI.A.

⁶This was also used by [Le Barbanchon et al. \[2021\]](#).

⁷The variety of features on both sides is summarized in Table C.3.

⁸As noted in [Algan et al. \[2020\]](#), this information is subject to measurement errors and limitations, such as that we do not observe when vacancies are filled through channels outside the PES. To limit this problem, we complement these data with extensive administrative data on all hires (*Déclarations préalable à l'embauche*) when we observe a hire within a firm with an identifiable job advertisement in the PES.

2 A first look at the recommendations generated by the two RSs

In this section, we undertake a comparative analysis of the two RS described in section 1: the ranking derived from the preference score $\mathcal{U}(i, j)$ (henceforth referred to as the \mathcal{U} -ranking) and the ranking derived from the ML-based score $S(i, j)$, which we post-process into a new score called $\mathcal{P}(i, j)$. As section 2.2 will show, this monotonic transformation of the ML score allows for interpretation as a “hiring probability”. Importantly, as the transformation is monotonic, the rankings derived from S and \mathcal{P} are identical for any given individual. We denote this ranking the \mathcal{P} -ranking throughout the analysis.

We define two recommendation sets for each job seeker: one based on the \mathcal{U} -ranking and the other on the \mathcal{P} -ranking. For either score, recommendations are made by selecting the k job vacancies with the highest value for each score.

2.1 Evaluating RSs predictive performance

Figure 1.2 shows the performance of different RSs that we considered when building our final ML-based RS. The figure on the right panel compares the performances in terms of recall@100 on the test set of different RS. Progressively including additional variables (such as previously considered vacancies) yields huge improvements on the recall.⁹ The first RS (“fixed weights”) corresponds to the \mathcal{U} -ranking, the preference-based RS inspired from the PES’s current one. As the graph shows, the recall@100 is very low, around 5%. The second recommendation system considered uses the same variables as those used to build the matching score, but instead of giving them fixed weights, it optimizes them to best predict the return to employment. This leads to improvements, but the recall@100 is still modest, remaining below 20%. The last two RS consider a broader set of variables. The first of the last two, based on neural networks, follows the method described in section 1.2 and is our \mathcal{P} -ranking. The second uses a different machine learning method based on ensembling and uses variables that explicitly describe the interactions between the variables characterizing the job supply and the job seekers (e.g. the distance between a job seeker and an establishment). Both RSs perform significantly better than the first two. The neural network achieves a recall@100 of about 57.5% and the last one an even higher recall@100. The disadvantage of the last system is its speed, especially when making recommendations. The neural network model takes about one hour to train and about 0.07 seconds to generate a set of recommendations for a given jobseeker; these figures are 2 hours and 10 seconds respectively for the last model.

⁹See the definition of the recall@k in section 1.1.

The right panel of Figure 1.2 shows how the recall of the last model varies with the number of recommendations. For 5 recommendations, the proportion is as large as almost 20%. As shown in the figure, the proportion increases progressively when the number of recommendations increases.

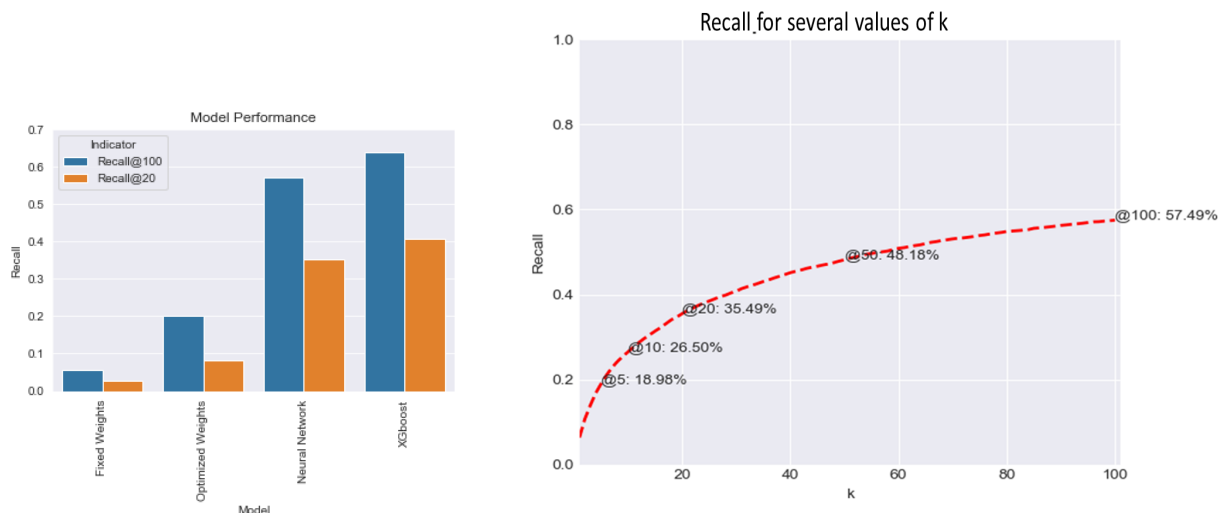


Figure 1.2: Performance on the test set of different RSs.

2.2 Estimation of the matching probability using the ML score as predictor

To facilitate further comparison of the two RSs, we post-process the ML score to transform it into an interpretable hiring probability. Drawing on the approach of Chernozhukov et al. [2018], we use a generic logistic predictor of the matching probability, given the ML algorithm's output score and based on observations of repeated applications and potential hires.

Let $M_{i,j}^* \in \{0,1\}$ be the latent variable that takes the value 1 if there is a match for a pair of job seeker-firm (i, j) after they meet. Let $C_{i,j} \in \{0,1\}$ be the observed variable that takes the value 1 if the job seeker i applies to the firm j 's vacancy. Importantly, after this process, the observed hiring dummy between i and j is $M_{i,j} = M_{i,j}^* C_{i,j}$.

We want to characterize the true probability of i being matched with j conditional on the very rich information $X_{i,j}$ available to us, namely $\mathbb{P}(M_{i,j}^* = 1 | X_{i,j})$. There are three main difficulties in estimating this true conditional probability. First, there is a selection problem, since we only observe matches conditional on a past interview $C_{i,j} = 1$, so the variable $M_{i,j} = M_{i,j}^* C_{i,j}$. Second, since we want to consider all potentially relevant covariates at our disposal, this is a high-dimensional setting. Third, ML algorithms, and

in particular those of section 1.2, generally do not produce a consistent estimator of the matching probabilities, but provide excellent predictive performance of future matching leveraging the complex interactions between the components of $X_{i,j}$. We provide a framework that allows to estimate the best predictions of the matching probabilities with a logistic predictor, given the score $S_{i,j}$ produced by the recommendation system.¹⁰

Denote by \mathcal{F}_j the sigma-algebra generated by the vector of past applications up to j , i.e., $\{C_{i,k} = 1, k = 1, \dots, j\}$, and negative observed results $\{M_{i,k} = 0, k = 1, \dots, j\}$. The selection problem translates into the fact that our data only allows us to identify $P(M_{i,j(i)} = 1 | X_{i,j(i)}, \mathcal{F}_{j(i)-1}, C_{i,j(i)} = 1)$ instead of $P(M_{i,j(i)}^* = 1 | X_{i,j(i)})$. To deal with this selection problem, we make the following assumption of conditional independence of the matching $\{M_{i,j}^*, j \in \mathcal{J}\}$ and application processes $\{C_{i,j}, j \in \mathcal{J}\}$.

Assumption 1 (Selection on observables and markovian property).

$$\Theta(X_{i,j(i)}) := P(M_{i,j(i)} = 1 | X_{i,j(i)}, \mathcal{F}_{j(i)-1}, C_{i,j(i)} = 1) = P(M_{i,j(i)}^* = 1 | X_{i,j(i)}).$$

Given how large the set of covariates we are starting from is, this assumption makes sense. Then, we take advantage of observing the chronologically ordered sequence for an individual $i_0, 1(i_0), 2(i_0), \dots, j^{max}(i_0)$ as a sequential search model and analyze it as a discrete duration model [see, e.g., Tutz et al., 2016], where the conditional hazard rate is $\Theta(X_{i,j(i)})$. Define the shortened notation for the score $S_{i,j(i)} := S_{i,j(i)}(X_{i,j(i)})$ and $r(i, j)$ the rank of the vacancy j in the application set. We define the best logistic predictor of this conditional probability given the score and this rank is $\Lambda(\alpha_{r(i,j(i))}^* + \beta^* S_{i,j(i)})$, where Λ is the usual logistic function and $(\alpha_{r(i,j(i))}^*, \beta^*)$ minimizes the Kullback Leibler Information Criterion (KLIC) with $\Theta(X_{i,j(i)})$, see White [1982].¹¹ The parameters of this best logistic predictor $(\alpha_{r(i,j(i))}^*, \beta^*)$ can be consistently estimated using conditional maximum likelihood estimation (MLE).

Estimation For the estimation we use the sequence $\{M_{i,j(i)}\}_{i=1, \dots, N; j=1, \dots, n(i)}$, where $n(i)$ is the number of observed applications for jobseeker i and N is the number of observed job seekers. Taking into account completed and censored spells [see, e.g., Tutz et al.,

¹⁰Note that the objective function (1.3) of this algorithm, whose purpose is to rank, is invariant to the addition of an individual specific effect α_i . However, this does not change the interpretation of our object of interest here, which is the best predictor *given the score and data used*. Of course, a different training could change the values of the estimated coefficients. Alternatively, one could use a logit with fixed effects approach similar to the one of section 3.5 to account for these potential shifts. However, this importantly limits the predictions to the set of “movers” (here 427 individuals) and even on them we typically observe few applications in this period (median of 3).

¹¹Thus, White [1982] also suggest this is a “minimum ignorance” solution. When the model is correctly specified, this identifies the true parameters.

2016, page 52], the estimation can be done using conditional MLE, considering the log-likelihood function, conditional on the scores produced by the RS, given by

$$\begin{aligned} \mathcal{L}(\alpha, \beta | M, S) = & \sum_{i=1}^N M_{i,n(i)} \ln(\Lambda(\alpha_{r(i,n(i))} + \beta S_{i,n(i)})) \\ & + \sum_{i=1}^N \sum_{j \in \mathcal{J}(i) \setminus \{n(i)\}} (1 - M_{i,j}) \ln(1 - \Lambda(\alpha_{r(i,j)} + \beta S_{i,j})). \end{aligned}$$

There is a simple generalization of the former expression to consider $r(i, j)$ the rank of vacancy j in the application set of job seeker i , but also $q(i, j)$ the rank of i in the applicant pool for job j . The likelihood expression in this case is written as

$$\begin{aligned} \mathcal{L}(\alpha, \beta | M, S) = & \sum_{(i,j): C_{i,j}=1} M_{i,j} \ln(\Lambda(\alpha_{q(i,j)}^v + \alpha_{r(i,j)}^{js} + \beta S_{i,j})) \\ & + \sum_{(i,j): C_{i,j}=1} (1 - M_{i,j}) \ln(1 - \Lambda(\alpha_{q(i,j)}^v + \alpha_{r(i,j)}^{js} + \beta S_{i,j})), \end{aligned}$$

where α^v and α^{js} are the sequences of “weariness” effects for vacancies and job seekers.

Estimation results Note that the potential for improving the value of the unemployed jobseeker is greater in this case than in the previous one. However, the estimation is performed on 34,255 randomly selected job seekers in the test set, representing 84,538 applications. As expected, the estimated coefficient of β of 0.061 is significantly positive at the 1% level. This result is robust to various specifications, including application and interview rank effects (0.038 and 0.047, respectively) (see table 1.2). Overall, this validates the content of the ML score $S_{i,j}$ in terms of its potential to reflect hiring chances. From now on, instead of $S_{i,j}$, we will think of our estimated best logistic predictor given the score in column (1) of table 1.2 as

$$\mathcal{P}(i, j) := \Lambda(0.061 S_{i,j} - 4.113) \tag{1.5}$$

which is our best prediction of the probability of hiring $p(i, j)$.

2.3 Contrasting preference-based and ML-based rankings

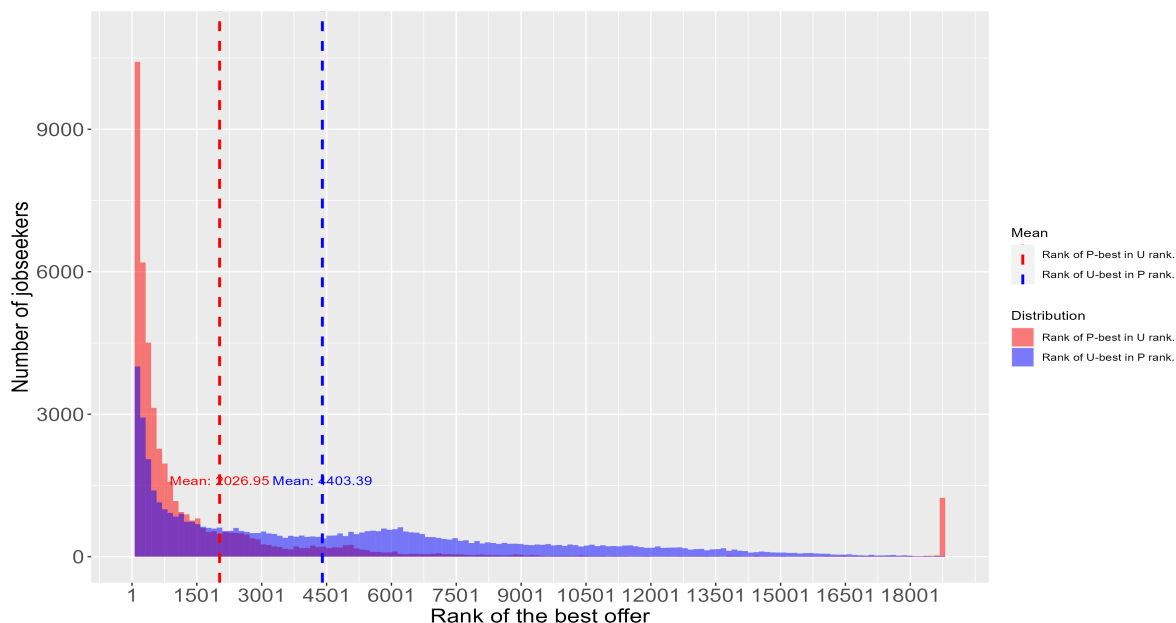
The two rankings are very different The two rankings are positively correlated for a large part of the population (median at 0.14 and first and third quartiles at 0.10 and 0.19 respectively), but there are still significant differences in the ranking of vacancies between the two RSs. To explore this further, we compare for each type of jobseeker i the optimal

Table 1.2: Estimates of the best logistic predictor given the ML score

Method	(1)	(2)	(3)
Score $S_{i,j}$	0.061*** (0.0029)	0.038*** (0.0030)	0.047*** (0.0030)
With application rank	No	Yes	Yes
With interview rank	No	No	Yes
Intercept	-4.113*** (0.0559)	-2.994*** (0.0570)	-2.538*** (0.0575)
AIC	28,040	25,116	23,897

Notes: On a half of the job seekers present in the test sample (weeks 44-48 of 2019): 79,097 applications, 3,469 matches, 34,255 job seekers. Significance levels: 1% : ***. "x applications" are dummies for the ranking of the application j in the list of applications of job seeker i . "x interviews" and "More than 11 inter." are dummies for the ranking of the candidate j in the list of recorded interviews for job ad j .

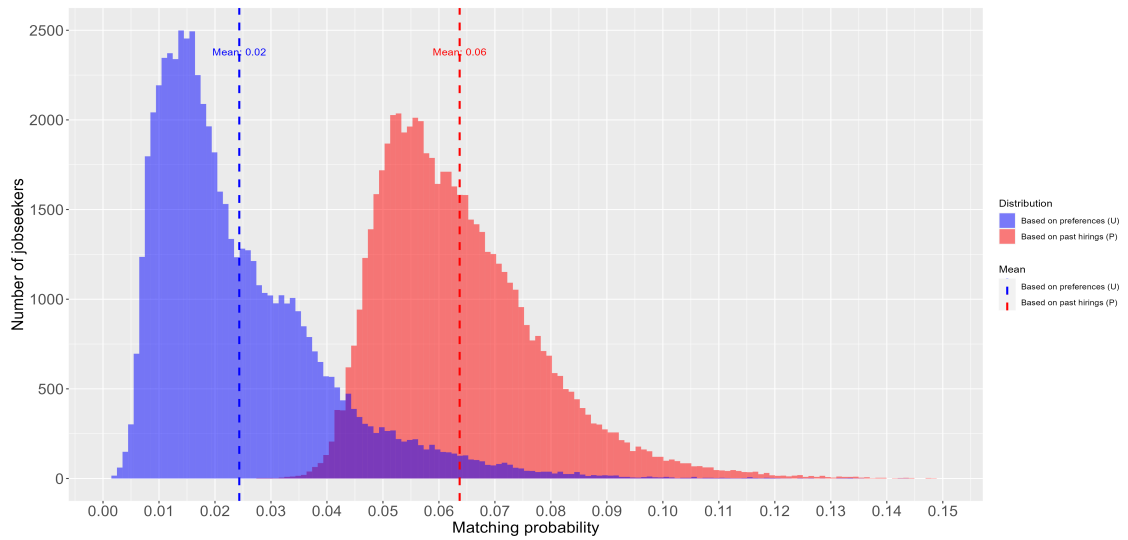
vacancy based on the \mathcal{P} ranking, denoted by $j^{\mathcal{P}}(i)$, and the optimal vacancy based on the \mathcal{U} ranking, denoted by $j^{\mathcal{U}}(i)$. We first compare the respective ranks of these optimal vacancies: the rank of $j^{\mathcal{P}}(i)$ in the \mathcal{U} -ranking: $r^{\mathcal{U}}(i, j^{\mathcal{P}}(i))$, and symmetrically the rank of $j^{\mathcal{U}}(i)$ in the \mathcal{P} -ranking: $r^{\mathcal{P}}(i, j^{\mathcal{U}}(i))$. Figure 1.3 shows the distribution of these ranks. While some individuals have optimal recommendations according to the two ranks that match, this is a small minority. For most, the ranks considered are very large. The median of $r^{\mathcal{U}}(i, j^{\mathcal{P}}(i))$ is 381 (top 2%) and that of $r^{\mathcal{P}}(i, j^{\mathcal{U}}(i))$ is 3,093 (top 16%).



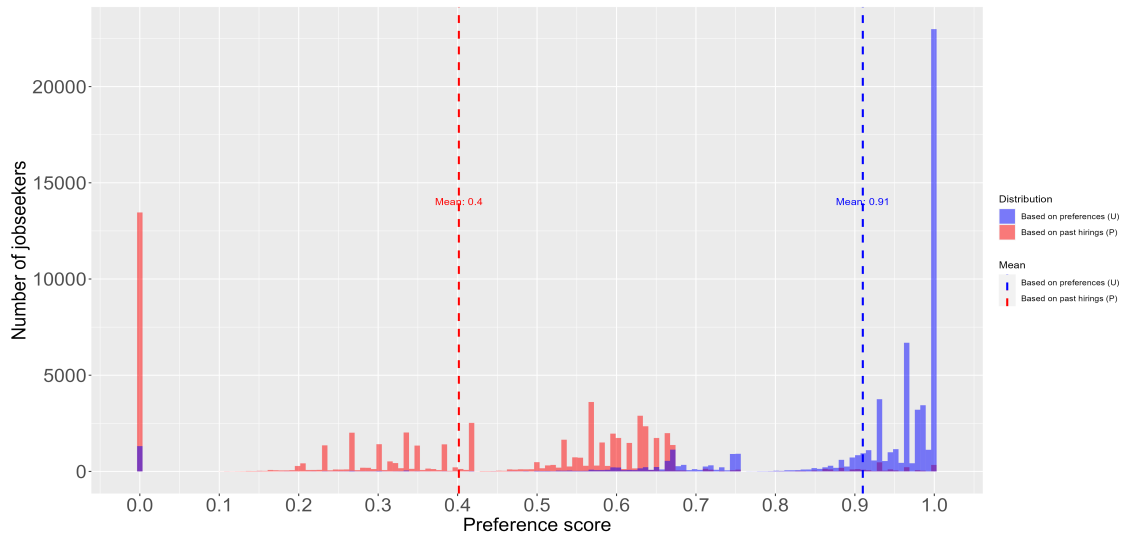
Notes: 60,299 job seekers whose main sector is transportation and logistic in the Rhône-Alpes region – ISO weeks 44-48 of 2019 – 18,873 vacancies available at that period in this sector. Distributions of the ranks of the best recommendations based on past hirings (\mathcal{P}) and elicited preferences (\mathcal{U}) in each other rankings. The small bunch at the right gathers vacancies associated with some best recommendations according to \mathcal{P} but ranked after 18,800 according to \mathcal{U} as they have a preference score of 0 and are ranked by distance to the job seekers.

Figure 1.3: Comparison of the rankings of the best recommendations wrt the other ranking

Differences between top- \mathcal{P} and top- \mathcal{U} job ads Thanks to the estimation result obtained in Section 2.2 we now give $j^{\mathcal{P}}(i)$ and $j^{\mathcal{U}}(i)$ a quantitative interpretation. Figure 1.4(a) shows the distribution of the hiring probabilities for the two vacancies: $\mathcal{P}(i, j^{\mathcal{P}}(i))$ and $\mathcal{P}(i, j^{\mathcal{U}}(i))$. The median value of the maximum hiring probability for each individual $\mathcal{P}(i, j^{\mathcal{P}}(i))$ is 0.06, sharply contrasting with the hiring probability for the optimal vacancy according to the adequacy criterion (0.02). Although the probability of hiring from the best vacancy in the \mathcal{P} -ranking is higher than the probability of hiring from the \mathcal{U} -ranking, it is worth noting that this probability in absolute terms is not so large. Even more pronounced differences arise in the matching scores $\mathcal{U}(i, j^{\mathcal{U}}(i))$ and $\mathcal{U}(i, j^{\mathcal{P}}(i))$. As shown in Figure 1.4(b), the distribution $\mathcal{U}(i, j^{\mathcal{U}}(i))$ has a substantial mass at 1 (median 0.98), indicating that for many job seekers there are vacancies that meet all their criteria. Conversely, for the optimal vacancy according to the hiring probability, there is a significant mass at zero (median 0.46).



(a) Distributions of hiring probabilities



(b) Distributions of preference score

Notes: 60,299 job seekers whose main sector is transportation and logistic in the Rhône-Alpes region – ISO weeks 44-48 of 2019 – 18,873 vacancies available at that period in this sector. *Upper panel:* Histograms of the hiring probabilities for the best recommendations in both systems. *Lower panel:* Histograms of the preference score for best recommendations in both systems.

Figure 1.4: Comparison of the best recommendations in the two rankings: hiring probabilities and preference score

2.4 Assessing the potential of combining the two RSs in the field

Empirically, we observed clear differences between rankings based on utility (\mathcal{U}) and those on hiring probability (\mathcal{P}). This observation raises the question of whether com-

binning these approaches could yield better outcomes in real-world settings. Building on this offline analysis, we conducted an exploratory field test¹² in March, 2022 to answer the two following questions: (1) how do job seekers perceive recommendations from different algorithmic sources in terms of fit and hiring chances? Do ML-based recommendations focused on hiring probabilities remain relevant even if they diverge from seekers' criteria? (2) can combining hiring-trained recommendations and proxies for job seekers' utility outperform either algorithm alone?

2.4.1 Experimental design

Treatment groups Job seekers were randomly assigned to one of five treatment groups, each corresponding to a different recommendation algorithm: \mathcal{U} -REC, \mathcal{P} -REC, MIX^{-1/4}, MIX^{-1/2}, and MIX^{-3/4}.

- **Preference-based recommendations (\mathcal{U} -REC):** This algorithm is inspired by the France Travail expert system and recommends jobs based on the fit between the job seeker's preferences and the job vacancy characteristics, reflecting the utility the job seeker would derive. It computes a weighted sum of various criteria¹³, using the proprietary France Travail weights¹⁴. This algorithm will be referred to as \mathcal{U} -REC in the rest of the analysis.
- **ML-based recommendations (\mathcal{P} -REC):** This algorithm recommends jobs based on predicted hiring chances, focusing on vacancies where the job seeker is more likely to succeed. It corresponds to the RS based on hiring predictions described in Section 1. This algorithm will be referred to as \mathcal{P} -REC in the rest of the analysis.
- **Combined recommendations (MIX^{-1/4}, MIX^{-1/2}, MIX^{-3/4}):** This approach selects job recommendations by balancing the input from the \mathcal{U} -REC algorithm and the \mathcal{P} -REC algorithm, in order to reflect the expected utility, as suggested by our earlier analysis. The algorithm, referred to as MIX, selects recommendations along the Pareto frontier between \mathcal{U} and \mathcal{P} , utilizing ordinal ranks rather than cardinal scores¹⁵. The process involves: i) selecting a set of ads highly ranked by \mathcal{U} -REC, \mathcal{P} -REC, or both; ii) narrowing down this set based on \mathcal{P} -REC's top-ranked ads; and iii) reordering the final selection by \mathcal{U} -REC scores. A key parameter is the share of ads discarded

¹²Approved by the Institutional Review Board of the Paris School of Economics (PSE) and registered at the AEA's Registry for RCTs (<https://doi.org/10.1257/rct.8998-1.3>).

¹³Criteria include working hours, reservation wage, geographic mobility, contract type, skills, diploma, languages, experience, and driver's license.

¹⁴Further details are available in Appendix A.1. Unlike the \mathcal{U} algorithm discussed in Section 1, this version includes a filtering mechanism for geographic distance.

¹⁵Detailed explanation in Appendix A.1.

in step ii), influencing the similarity of MIX to either \mathcal{U} -REC or \mathcal{P} -REC. We consider three versions: $\text{MIX}^{-1/4}$, $\text{MIX}^{-1/2}$, and $\text{MIX}^{-3/4}$, which discard one-quarter, one-half, and three-quarters of the initial set, respectively. Thus, $\text{MIX}^{-1/4}$ is closer to pure \mathcal{P} -REC recommendations, while $\text{MIX}^{-3/4}$ aligns more with \mathcal{U} -REC. This approach will be referred to as MIX in the rest of the analysis.

The recommendation policies based on \mathcal{U} -REC and \mathcal{P} -REC significantly differ: the top-1 job ad recommended by \mathcal{P} -REC is included in the top-10 recommendations of \mathcal{U} -REC for only circa 15% of the job seekers; it does not appear among the top-100 recommendations of \mathcal{U} -REC for circa 64% of the job seekers. Figure A.1 (in Appendix) also shows the significant variations in \mathcal{P} and \mathcal{U} scores between the algorithms considered. Additional treatment arms provided complementary information to job seekers about the algorithms and job ad rankings, but this section focuses on the core comparison between the different recommendation criteria.

Survey protocol and data description The eligible population consists of job seekers registered at France Travail in the Auvergne-Rhône-Alpes region who were actively seeking employment. The randomization across the five treatment groups was successful (see Appendix Table A.1).

Out of the 50,495 individuals invited, 17.7% participated in the experiment, which involved completing an online survey. Participants were asked to rate two job ads (the top two recommended by their assigned algorithm) on relevance, fit with their preferences, and perceived hiring chances, followed by access to the top-10 list of job recommendations. More details about the survey protocol are included in to the Appendix A.3.

For the analysis, we focus on the subset of participants who completed the survey, as there is no evidence of differential attrition in survey completion rates (see Appendix Table A.2). Our analysis sample provides detailed individual-level data, including sociodemographic characteristics, job search criteria, resume details, and unemployment history. The outcomes analyzed include job seekers' ratings of ads (relevance, fit, hiring chances) and an engagement metric measured by clicks on recommended ads.

2.4.2 Experimental results

We estimate the following model at the job seeker-vacancy pair level using ordinary least squares (OLS):

$$Y_{ij} = \alpha + \sum_k \beta_k \{T_i = k\} + \gamma Z_i + \epsilon_{ij} \quad (1.6)$$

where T_i denotes the treatment assigned to the job seeker i , and Y_{ij} represents either the ratings of the top-2 job ads (evaluated in terms of overall rating, hiring probability, and fit to job seekers' search criteria) or interaction metrics (the number of clicks and the probability of clicking on at least one ad among the top-10 recommendations). The vector Z_i includes control variables reflecting the stratified randomization design. The coefficients β_k represent the Average Treatment Effect (ATE) of each algorithm relative to the \mathcal{U} -REC algorithm, which serves as the reference category. Standard errors are clustered at the job seeker level, and strata fixed effects are included in all specifications to control for baseline differences across strata.

Table 1.3 summarizes the results. First, we find that overall ratings do not significantly vary across algorithms. Although the coefficients for alternative algorithms to the \mathcal{U} -REC are positive, they are not statistically significant, indicating no notable improvement in job seekers' evaluation of recommended ads relative to the \mathcal{U} -REC baseline. This suggests that job seekers perceive no substantial difference in overall recommendation quality across the tested algorithms. To investigate the prevalence of negative reactions towards recommendations, we examine whether job seekers rated at least one of the top-2 ads as 1/10 or below 5/10. None of the algorithmic variants, including \mathcal{P} -REC and the MIX models, show a significant increase in the rate of "poor" job ads compared to the \mathcal{U} -REC baseline. These findings imply that the acceptability of recommendations from \mathcal{U} -REC and its variants is relatively similar: job seekers do not systematically view the alternative algorithms as delivering less acceptable recommendations.

We next consider job seekers' perception of how well the top-2 ads match their search criteria. The evidence here is mixed. For the \mathcal{P} -REC algorithm, the coefficient is 0.002, indicating a negligible difference from the control mean of 3.277. In contrast, the MIX^{-1/4} algorithm shows a more substantial increase, with a coefficient of 0.178, representing an approximate 5.4% rise relative to the control mean. This effect is statistically significant at the 5% level, suggesting that MIX^{-1/4} better aligns with job seekers' criteria than the \mathcal{U} -REC algorithm. While the MIX^{-1/2} algorithm also has a positive coefficient (0.119), it does not reach statistical significance. These results indicate that certain MIX model variants may slightly enhance the perceived fit of recommendations, though the improvements are not consistent across all algorithms. Turning to the hiring probability rating, which reflects job seekers' assessment of the ads' likelihood to lead to successful employment, both \mathcal{P} -REC and its closest MIX variants yield positive and statistically significant coefficients. For instance, the coefficient for \mathcal{P} -REC is 0.219, translating to a 6.5% increase over the control mean of 3.357. The MIX^{-1/4} and MIX^{-1/2} algorithms show similar positive and significant coefficients (0.254 and 0.255, respectively), indicating an approximately 7.6% increase over the control mean. These findings suggest that job seekers perceive recom-

recommendations from \mathcal{P} -REC and the MIX models as more likely to result in employment than those from the \mathcal{U} -REC algorithm.

Lastly, we examine interaction metrics, specifically whether job seekers clicked on at least one of the top-2 ads (Column 6 in Table 1.3) and the total number of clicks among the top-10 recommendations (Column 7). In Column 6, the coefficients for all algorithms are near zero and not statistically significant, indicating no meaningful difference in the likelihood of clicking on the top-2 ads across treatments. This suggests that initial engagement remains relatively constant regardless of the algorithm used. However, Column 7 reveals a more nuanced pattern. While most algorithms have positive coefficients, only \mathcal{P} -REC MIX^{-1/4} and MIX^{-1/2} demonstrate statistically significant increases over the \mathcal{U} -REC baseline. The coefficient for \mathcal{P} -REC is 0.006, corresponding to a 14.6% increase over the control mean of 0.041 clicks per person, and is significant at the 5% level. The coefficient for MIX^{-1/4} is 0.005, corresponding to a 12.1% increase over the control mean (significant at the 10% level). The MIX^{-1/2} algorithm exhibits an even larger effect, with a coefficient of 0.008, marking an 18% increase over the control mean, significant at the 5% level. These results indicate that, although job seekers do not necessarily click on the top-2 ads more often, certain algorithms, particularly MIX^{-1/2}, drive higher overall engagement with job recommendations, as reflected in the increased clicks among the top-10 ads. This suggests limitations of \mathcal{U} -REC: while effective in identifying relevant ads when they exist, its performance declines when balancing multiple criteria, especially under a rigid weighting scheme. In contrast, the improvements seen with \mathcal{P} -REC and the MIX models indicate these algorithms may better navigate such trade-offs, leading to more engaging recommendations.

2.5 Overall takeaway

Our analysis reveals a significant divergence between the recommendations generated by the machine learning-based system focused on hiring probabilities (\mathcal{P}) and those from the preference-based system grounded in job seekers' search criteria (\mathcal{U}). To further investigate this discrepancy, we conducted a field evaluation of job seekers' satisfaction with, and adoption of, the ML recommender system, the preference-based system, and hybrid models integrating both approaches. The analysis led to several key observations. The ML system's recommendations were as acceptable to job seekers as those from the utility-based system. Importantly, its suggestions led to perceived chances of hiring, without increasing negative reactions. However, the most policy-relevant outcome may be the impact on click-through rates, as clicks reveal individuals' true preferences and are correlated with applications and, ultimately, returns to employment. The hybrid vari-

Table 1.3: Treatment differences in recommendations appreciation

Dependent variable	Top-2 ads				Top-10 ads		
	Overall rating	Overall rating 1/10	Overall rating < 5/10	Fit to search criteria	Hiring	Clicked on ad	Clicked on ad
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
\mathcal{P} -REC	0.011 (0.085)	0.001 (0.005)	0.006 (0.014)	0.002 (0.089)	0.219 (0.088)**	-0.001 (0.005)	0.006 (0.003)**
MIX- ¹ / ₄	0.051 (0.084)	-0.001 (0.005)	-0.003 (0.014)	0.178 (0.090)**	0.254 (0.087)***	-0.001 (0.005)	0.005 (0.003)*
MIX- ¹ / ₂	0.035 (0.082)	-0.001 (0.005)	0.001 (0.014)	0.119 (0.087)	0.255 (0.086)***	0.000 (0.005)	0.008 (0.003)**
MIX- ³ / ₄	0.016 (0.085)	0.002 (0.005)	0.008 (0.014)	0.004 (0.090)	0.081 (0.088)	-0.006 (0.005)	0.001 (0.003)
Strata fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N. Obs.	17 842	17 842	17 842	17 842	17 842	17 842	89 210
N. Clusters	8921	8921	8921	8921	8921	8921	8921
Control mean (\mathcal{U} -REC)	5.163	0.039	0.361	3.277	3.357	0.041	0.041

Note: The \mathcal{U} -REC treatment group is used as the reference category. Outcomes (1), (4), and (5) are ratings on a scale of 1 to 10. Outcomes (2), (3), (6), and (7) are binary variables. Standard errors are clustered at the job seeker level. Robust standard errors are in parentheses. *, **, ***: significance at 10%, 5%, and 1%.

ants (MIX-¹/₄ and MIX-¹/₂) proved to be the most effective algorithms in driving higher click-through rates, suggesting that better balancing hiring probabilities and job seekers' preferences, results in more engaging and impactful recommendations.

3 Job search model with RSs

Our findings suggest the necessity of an integrated approach to design recommendation systems that effectively serve job seekers' interests. To address this challenge, we develop a theoretical job search model that incorporates both the utility derived from job characteristics and the probability of a successful application. This model provides a comprehensive framework for understanding job seekers' decision-making processes and for designing optimal recommendation systems that enhance their search outcomes.

We propose that job seekers consider two key factors in their search: the utility U associated with a job and the probability P of a successful application. Higher utility U indicates that the job has characteristics valued by the job seeker. Additionally, as in standard job search models, there is a reservation utility U^* . We posit that the recommendation system based on search criteria (\mathcal{U}), signals $U - U^*$ and thus ranks jobs by utility. Conversely, the second system, \mathcal{P} , reflects P and ranks jobs by the chances of success.

We acknowledge that \mathcal{U} and \mathcal{P} may not exactly correspond to $U - U^*$ and P , respectively. They are likely functions of both $U - U^*$ and P , i.e., $\mathcal{U} = \mathcal{U}(U - U^*, P)$ and $\mathcal{P} = \mathcal{P}(U - U^*, P)$. Therefore, even if \mathcal{U} primarily reflects utility, it may also contain information about the probability of success, and vice versa. We will revisit this point later.

3.1 Model framework

We develop a model in which the application behavior and performance in the job search process is based on the two quantities U and P . We consider job seekers of different types x facing a set of vacancies of different type y . Job seekers can allocate effort in searching for jobs either by themselves ($d = 0$) or through the use of the platform ($d = 1$). We denote this decision $s_d \in \{0, 1\}$ and for simplicity consider the case where they only use one channel $s_0 + s_1 = 1$.¹⁶

Following the previous discussion, for a jobseeker i with characteristics x , the heterogeneity of vacancies y refers to differences in two dimensions:

- The first is the utility associated with holding the job in question, $U(x, y) + \varepsilon_{i,y}$. Jobs differ in a number of parameters - salary, location, type of contract, working hours, type of occupation, skills required, etc. - that provide different levels of utility to the job seeker. The job seeker has access to both $U(x, y)$ and $\varepsilon_{i,y}$. However, the researcher only has access to $U(x, y)$. We assume $\varepsilon_{i,y}$ is distributed as a logistic distribution with scale parameter σ , with $F(z/\sigma)$ its cumulative distribution and $F(\cdot)$ the standard logistic function. A large σ means that U is a limited signal on the utility associated with the job.
- The second dimension is the perceived probability of success of an application $\pi(x, y)$. This uncertainty about the success of an application corresponds to the competition from other job seekers, but also to the uncertainty about the suitability of the profile of the jobseeker with characteristics x for the profile sought by the recruiter for the job y . This information about mutual suitability is revealed downstream during the hiring process, but is not known at the time of the application. While job seekers expect that their application will result in a hiring with a subjective probability $\pi(x, y)$, there is a true probability that this will happen $p(x, y)$. We further assume that the potential deviation from rationality is deterministic, i.e., there exists a function Π such that $\pi(x, y) = \Pi(p(x, y))$.

¹⁶Similarly to Van den Berg and Van der Klaauw [2006], we can also introduce search effort in the two channels s_0, s_1 and costs of search here $c(s_0, s_1) = (s_0^\gamma + s_1^\gamma)^{2/\gamma}$, with $1 < \gamma < 2$.

For a jobseeker i with characteristics x , a vacancy y is a lottery with rewards $U(x, y) + \varepsilon_{i,y}$ in case of success, and real and perceived chances of success $p(x, y)$ and $\pi(x, y)$.

Job matches are destroyed at the exogenous rate q . While unemployed, workers receive the flow utility $u(b)$. If they search for a job through channel d , job seekers with characteristic x have a discounted present value of utility $V_{d,u}$. If they accept a job with characteristic y , the discounted present value of their utility is $V_e(x, y, \varepsilon_{i,y})$.

If a jobseeker of type x sees a job posting with characteristics y , he or she will decide to apply (noted as $C(x, y, \varepsilon) = 1$), expecting to be hired with subjective probability $\pi(x, y)$. This decision to apply is made as long as the discounted expected benefits of applying is higher than the discounted value of remaining unemployed ($V_{d,u}(x)$) and continuing to search. The discounted expected benefit of applying writes as $\pi(x, y)V_e(x, y, \varepsilon_{i,y}) + (1 - \pi(x, y))(V_{d,u}(x) - R) - k$. In this expression, there are two costs: k is the cost of applying and R refers to the (psychological) costs of being rejected.

When job seekers search for jobs on their own ($s_0 = 1$), they draw vacancies from a distribution $F_0(y)$. The RS also draws job vacancies from the same pool of vacancies F_0 , but makes for each jobseeker x an individualized selection $S(x, y)$, resulting in a distribution $F_1(x, y)$. Job seekers will find jobs with arrival rate α_0 if they search on their own or α_1 if they search using the recommendation system. We assume that job seekers have consistent expectations $\Pi(p(x, y)) = p(x, y)$ to develop the model and study the impact of RSs. While the case in which job seekers hold inconsistent perceptions of their chances of success ($\Pi(p(x, y)) \neq p(x, y)$) is relevant, a detailed discussion of this issue falls outside the scope of this chapter.

3.2 Searching in absence of the RS

We first analyze the application behavior of job seekers in the absence of a RS. This corresponds to $s_0 = 1$ and $s_1 = 0$. We introduce

$$U_0^*(x, p) = rV_{0,u}(x) - \bar{R} + \frac{\bar{k} + \bar{R}}{p}, \quad (1.7)$$

which from the following proposition we interpret as the reservation utility for jobs that job seekers expect to get with probability p , and where hereafter we rescale R and k to $\bar{R} = R(r + q)$ and $\bar{k} = (r + q)k$.¹⁷ The search behavior is summarized by Proposition 3.1.

¹⁷To follow-up on the preamble of this section, U is seen as a signal on $U(x, y) - U_0^*(x, 1)$.

Proposition 3.1. *The discounted value $V_{0,u}(x)$ for an unemployed jobseeker of type x absent the RS, is the solution of the equation*

$$rV_{0,u}(x) = u(b) + \frac{\alpha_0}{r+q} \int_{C(x,y,\varepsilon_y)=1} p(x,y) (U(x,y) - U_0^*(x,p(x,y)) + \varepsilon_y) dF_0(y) dF(\varepsilon_y).$$

Job seekers of type x decide to apply on vacancies of type y characterized by a utility $U(x,y)$ and a perceived probability of success $\pi(x,y)$ if

$$C(x,y,\varepsilon_y) = 1 \Leftrightarrow U(x,y) - U_0^*(x,p(x,y)) > -\varepsilon_y. \quad (1.8)$$

Proof. See model appendix B. □

A useful reparametrisation. In this setup, only two quantities summarize the interest for jobs of characteristics y : 1) the utility for the job $U(x,y)$, 2) the probability to be hired on the job $p(x,y)$. Thus, we can reparametrize the distributions of the characteristics of the vacancies in this twodimensional space and use $F_0(p,u)$ hereafter. From now on, we will omit the x characteristic to simplify notations, where this does not lead to ambiguity. For example, we write $V_{0,u}$ instead of $V_{0,u}(x)$ and $U_0^*(p)$ instead of $U_0^*(x,p)$.

The decision to apply is based on the “surplus” defined as $\Delta(u,p) := u - U_0^*(p)$, and amounts to:

$$\Delta(u,p) + \varepsilon_y > 0. \quad (1.9)$$

Note that, given equation (1.7), $U_0^*(p) = U_0^*(1) - (\bar{k} + \bar{R}) + (\bar{k} + \bar{R})/p$. Thus, $\Delta(u,p)$ also writes as:¹⁸

$$\Delta(u,p) = (u - U_0^*(1)) + \bar{k} + \bar{R} - \frac{\bar{k} + \bar{R}}{p}. \quad (1.10)$$

Using Proposition 3.1, the discounted value for an unemployed jobseeker searching through the standard search channel takes the simple form:

$$rV_{0,u} = u(b) + \frac{\alpha_0}{r+q} E(\Gamma(P,U)), \quad (1.11)$$

where

$$\Gamma(p,u) := pP_\varepsilon(\Delta(u,p) + \varepsilon > 0)E_\varepsilon(\Delta(u,p) + \varepsilon | \Delta(u,p) + \varepsilon > 0). \quad (1.12)$$

¹⁸Given this expression, a better notation would be $\Delta(u,p,U_0^*(1))$. We don't use it however, so as to keep the notation simple. We will come back to this issue when discussing optimal RSs and their potential impact on reservation utility, i.e. a change in $U_0^*(1)$.

The distributional assumption on ε implies:¹⁹

$$\Gamma(p, u) = p\sigma \log(1 + e^{\Delta(u,p)/\sigma}). \quad (1.13)$$

Note that in the expression (1.12), $\Gamma(p, u)$ can be decomposed as the product of the probability to apply, $P_\varepsilon(\Delta(u, p) + \varepsilon > 0) = F(\Delta(u, p)/\sigma)$; the probability of being hired conditional on applying, i.e. p and a third term which is the expected “surplus” utility on the jobs selected for application. Note also, that when $\sigma \rightarrow 0$, $\Gamma(p, u) \rightarrow p\Delta(u, p)1(\Delta(u, p) > 0)$. Thus, when the signal on the utility of the job is good, then the expected value for a given job for the job seeker is simply the product of the chances to be hired p and the surplus $\Delta(u, p)$, provided this surplus is positive.

3.3 Identification of Γ

Before deriving the optimal RS and the corresponding index, it is useful to examine the identification of Γ . As equation (1.13) shows, such an identification relies on that of $\Delta(u, p)$. And, as shown by equation (1.9), such a quantity is easily identified from the observation of applications conditional on utility and matching probability using the functional form of equation (1.10). We denote by $C_{i,j} = 1$ if job seeker i applied to job posting j , and 0 otherwise.

Proposition 3.2. *Assuming ε is distributed as a logistic variable with scale parameter σ ; assuming sequences of individual i application decisions on vacancies j are observed and, following equation (1.9) and (1.10), identify the α , β , and γ parameters:*

$$P(C_{i,j} = 1 | P_{i,j}, U_{i,j} - U_{0,i}^*(1)) = F(\alpha(U_{i,j} - U_{0,i}^*(1)) - \beta/P_{i,j} + \gamma) \quad (1.14)$$

then σ , $\bar{k} + \bar{R}$, $\Delta(u, p)$ and $\Gamma(p, u)$ are identified:

- σ is identified as $1/\alpha$;
- $\bar{k} + \bar{R}$ is identified as β/α or γ ;
- $\Delta(u, p)$ is identified as $(u - U_0^*(1)) + \gamma/\alpha - (\beta/\alpha) \times 1/p$;²⁰
- $\Gamma(p, u)$ is identified as $p \log(1 + e^{\alpha(u - U_0^*(1)) + \gamma - \beta/p})/\alpha$.

As already mentioned, the \mathcal{U} score can be considered as a signal on $U_{i,j} - U_{0,i}^*(1)$. In the empirical application, we introduce it directly.

¹⁹More generally, relaxing the distributional assumption on ε , $\Gamma(p, u)$ can be written as $\Gamma(p, u) = p\sigma K(\Delta(u, p)/\sigma)$, with K an increasing function such that $\lim_{x \rightarrow \infty} K(x)/x = 1$ and $\lim_{x \rightarrow -\infty} K(x) = 0$.

²⁰As discussed in footnote 18, a more precise notation would also include $U_0^*(1)$ argument.

3.4 Search using a RS

A RS operates in two directions:

1. It selects vacancies from the initial distribution associated with an index $S(p, u)$. Here we consider a selection of vacancies that are above the quantile of order $1 - s$, for a given s . The selection rule is thus written as

$$dF_1(p, u) = \frac{1\{F_S(S(p, u)) > (1 - s)\}}{s} dF_0(p, u), \quad (1.15)$$

where F_S is the cumulative distribution if the index S .

2. It exposes job seekers to a more intense flow of job vacancies: instead of α_0 , job seekers are exposed to a flow of opportunities with intensity α_1 . Although we start the discussion with $\alpha_1 = \alpha_0$, we will also consider $\alpha_1 > \alpha_0$, as well as the case where α_1 depends on s : $\alpha_1(s)$, and in particular the situation where it increases with s .

We assume that job seekers are myopic: they do not adjust their reservation utility following the exposition to opportunities through the RS. In such a case, the discounted value $rV_{1,u}^m$ for an unemployed job seeker exposed to recommendations is written as follows:

$$rV_{1,u}^m(S, s, \alpha_1) = u(b) + \frac{\alpha_1}{r + q} E(\Gamma^m(P, U) | F_S(S(P, U)) > 1 - s). \quad (1.16)$$

The function Γ^m is the same as Γ in the case job seekers are searching on their own and we label it as m to emphasize myopic job seekers.²¹

Defining the best RS in the case of myopic job seekers is simple. The result is summarized in the following proposition:

Proposition 3.3. *For a RS based on an index $S(p, u)$ selecting for a job seeker of type x the top $s\%$ of vacancies ranked according to S , the discounted value for a myopic unemployed jobseeker writes as (1.16). Moreover:*

- *The best RS is based on the index $S(p, u) = \Gamma^m(p, u)$;*
- *When $\alpha_1 \geq \alpha_0$, a sufficient condition for an S based RS to improve the value of a myopic unemployed jobseeker compared to searching through the standard channel is that $E(\Gamma^m(P, U) | S = z)$ increases in z .*

²¹Recall that from equation (1.13) that $\Gamma(p, u) = p\sigma \log(1 + e^{\Delta(u,p)/\sigma})$ with $\Delta(u, p) = u - U_0^*(p)$, and, as shown in equation (1.7), $U_0^*(p) = rV_{0,u} - \bar{R} + (\bar{k} + \bar{R})/p$. Myopic job seekers have a discounted value given by equation (1.16), but in which Γ is still defined using the discounted value absent the RS.

- Following the previous point, for $\alpha_1 \geq \alpha_0$, the RS based on the index $S(p, u) = \Gamma^m(p, u)$ always improves the discounted value for a myopic unemployed jobseeker compared to searching through the standard channel.

Proof. See model appendix B.1. □

The result clearly states that the RS based on $\Gamma^m(P, U)$ dominates the RSs based on either only P or only U , or the RS based on $\Delta(U, P)$, the index which is driving applications, as shown by the equation (1.9). The result also shows that it is not guaranteed that these latter recommendation systems improve the standard search, especially when the two indices P and U are negatively correlated at the top of the distribution of P or U .

In our empirical analysis in section 4, we will thus compute the quantities

$$E(\Gamma^m(P, U) | F_S(S(P, U))) \geq 1 - s$$

for $S(p, u) = \Gamma^m(p, u)$, $S(p, u) = p$, $S(p, u) = u$ and $S(p, u) = \Delta(u, p)$ for each individual in our sample and different values of s . This will help to assess the gains associated with using the optimal RS and the magnitude of the gains compared to the two intuitive and competing RSs based on p and u , as well as whether these two RSs actually improve the jobseeker's situation.

So far we have discussed the changes associated with the introduction of a RS that maintains the frequency of exposure of job seekers to vacancies $\alpha_1 \geq \alpha_0$. Indeed, RSs can be associated with a change in α_1 . On the one hand, we could consider that RSs extract relevant vacancies at low cost and thus expose job seekers to more relevant vacancies, i.e. $\alpha_1 \geq \alpha_0$. The benefits of using the recommendation system based on $\Gamma^m(P, U)$ obviously still apply when $\alpha_1 \geq \alpha_0$. On the other hand, we might consider that if the RS is very selective and only focuses on the very top of the job listings (s small), a large α_1 might not be sustainable. So to some extent there is a function $\alpha_1(s)$ that might decrease with s . However, as long as

$$\alpha_1(s) E(\Gamma^m(P, U) | F_{\Gamma^m}(\Gamma^m(P, U))) \geq 1 - s \geq \alpha_0 E(\Gamma^m(P, U)),$$

the RS improves the situation of job seekers. This means that there could be a value of s below which the gains in the quality of job vacancies are outweighed by the scarcity of possible recommendations.

3.5 Empirical validation of the model

Using the calibrated hiring probabilities estimated from Equation 1.5, as well as the data on applications and clicks on job postings described in Section 1.4, we now estimate the model developed above. This allows us to estimate job seekers' preferences for the job characteristics and finally to propose an interpretation of the two RSs introduced in Section 1.

The model shows how job seekers' preferences for the different characteristics of jobs affect application decisions. In this section, we use observations on job seekers' applications to estimate these preferences. As in Hitsch et al. [2010]²² or Le Barbanchon et al. [2021], given the threshold based decision rule in equation (1.8), these preferences can be estimated using a discrete choice model.

We consider the set of vacancies on which job-seekers have clicked and estimate a logit model of application decisions following equation (1.14) in proposition 3.2, the decision to apply depends on $U_{i,j} - U_{0,i}^*(1)$ and $1/P(i, j)$ that we proxy using $\mathcal{U}(i, j)$ and $1/\mathcal{P}(i, j)$.

The estimates we present in this section have two objectives. The first is to provide an empirical validation of the model. We want to provide evidence that, other things being equal, when job seekers are exposed to vacancies with different levels of utility and different probabilities of success, their decision to apply integrates these two variables according to the predictions of the equation 1.9. To go as far as possible in identifying a causal relationship, we estimate the previous model with fixed effects.²³ In this respect, when considering the design of an RS, rejecting $\alpha = 0$ and $\beta = 0$ provides evidence that the optimal RS must be based on both \mathcal{U} and \mathcal{P} . The second objective is to identify a score that best approximates the quantity $\Delta(u, p)$ with the aim of constructing the optimal recommendation score. For this reason, we also examine specifications that do not involve individual effects.²⁴

The model we estimate takes thus the general form of a binary choice model with or without fixed effects:

$$\mathbb{P}(C_{i,j} = 1 | \text{click}_{i,j} = 1, W_{i,j}[\gamma_i]) = \Lambda \left(\alpha \mathcal{U}(i, j) - \frac{\beta}{\mathcal{P}(i, j)} \quad [+ \gamma_i] \right), \quad (1.17)$$

where $W_{i,j} := (\mathcal{U}(i, j), \mathcal{P}(i, j), X_i, Y_j)$. Leveraging the declared data at the PES, we use a

²²See for example their equation (9).

²³It is worth to note that strictly speaking the the model does not require the introduction of such fixed effects as the index \mathcal{U} is a measure of the difference between utility and reservation utility

²⁴In this respect, the estimation without fixed effects can be seen as a first step of a prediction task of the application decision based on a flexible form of \mathcal{P} and \mathcal{U} and even the components entering \mathcal{U} .

simple specification to estimate the preferences for jobs amenities $\mathcal{U}(i, j)$:

$$\mathcal{U}(i, j) = \sum_{k=1}^K w_k c_k(i, j), \quad (1.18)$$

where $\{w_k\}_{k=1, \dots, K}$ denotes a set of weights for each characteristic k and $c_k(i, j) \in [0, 1]$ represents a consistency criterion. This criterion reflects whether the k -th characteristic of the job seeker’s profile aligns with that of the vacancy. For example, the criterion for “reservation wage” takes the value of 1 if the wage offered in the vacancy exceeds the job seeker’s reservation wage. Similarly, for “geographic mobility”, it assumes the value of 1 if the job location’s distance falls within the job seeker’s maximum commuting range. This model is particularly advantageous as it enables comparison with the recommendation strategy employed by PES, wherein a normalized set of weights is uniformly applied across the entire population. Specifically, these weights are defined as follows: Occupation (0.332), Skills in occupation (0.332), Geographic mobility (0.997), Reservation wage (0.066), Diploma (0.033), Working hours (0.033), Driving license (0.033), Languages (0.033), Years of experience in occupation (0.033), and Duration and type of contract (0.003).²⁵ These weights constitute our baseline specification for $\mathcal{U}(i, j)$. However, for robustness, we also examine another specification using the components of \mathcal{U} instead of \mathcal{U} itself:

$$\mathbb{P}(C_{i,j} = 1 | \text{click}_{i,j} = 1, W_{i,j}[, \gamma_i]) = \Lambda \left(\alpha \sum_{k=1}^K \delta_k c_k(i, j) - \frac{\beta}{\mathcal{P}(i, j)} \quad [+ \gamma_i] \right), \quad (1.19)$$

where $W_{i,j} := (c_k(i, j), \mathcal{P}(i, j), X_i, Y_j)$. While we also consider the weights δ_k as given (either uniform or those used within the PES – see section 1.3) and estimate α and β as previously, we also let the weights unconstrained, and estimate the products $\alpha \delta_k$. This is in spirit close to [Hitsch et al. \[2010\]](#), [Chen et al. \[2023\]](#), where they estimate similar equations in the context of the marriage market.

Results appear in Table 1.4. The first and second columns (Logit) and (FE-Logit) present the results respectively with and without fixed effects, while the third column (FE-Logit unconstr.) presents the results of the specification (1.19) with unconstrained weights. For each of the three estimates, the variable $-1/\mathcal{P}(i, j)$ has a significant positive coefficient (which implies, as expected, that the probability of applying increases with \mathcal{P}). Moreover, these coefficients are very similar: 0.018 for the logit, 0.028 for the logit FE, and 0.026 for the logit FE with estimated weights. Similarly, the utility score coefficient \mathcal{U} has

²⁵[Field et al. \[2023\]](#) similarly utilize this specification and set of characteristics to assess vacancy value from the perspective of job seekers; see their Table 3 for further details.

Table 1.4: Estimates of the model of application on job postings

	(1) Logit		(2) FE-Logit		(3) FE-Logit unconstr.	
	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Utility score $\mathcal{U}(i, j)$ (α)	0.992**	0.194	1.101***	0.155		
Occupation					0.582***	0.104
Skills					0.175*	0.114
Reservation wage					0.236***	0.082
Languages					-0.010	0.229
Experience in occ.					-1.017***	0.339
Diploma					0.288**	0.118
Driving license					0.106	0.097
Geographic mobility					0.625***	0.214
Duration					0.139	0.068
Type of contract					0.015	0.004
- Inverse of $\mathcal{P}(i, j)$ (β)	0.018**	0.007	0.028***	0.004	0.026***	0.004
Intercept / Avg. indiv. FE	-1.908	0.179	-1.388	0.047	-1.372	0.04
Nb. observations	17,865		17,865		17,865	

Estimation of equation (1.19) modeling applications as a logit model with or without fixed effects and with weights constrained or unconstrained.

Notes: Our sample is the set of all applications for job seekers in the transportation and logistic sector during week 44 of 2019, leading to a hiring or not. Fixed effect estimation keeps 70,557 observations for 8,105 job seekers, and 869 of them applying at least once, and 757 "Movers" with one application and 1 click without application. Thus, 17,865 observations are kept for estimation. Estimation of results for a logit panel without fixed effects on the full sample are available in Appendix. Results are robust to the different negative sampling strategies we considered. Significance levels: < 1% : ***, < 5% : **, < 10% : *.

a positive and significant coefficient in each of the first two specifications and also very close, respectively in column order 0.992 and 1.101. As stressed above, the result in the second column is especially important as it validates the interpretation we sketched at the beginning of section 3, of the score $\mathcal{U}(i, j)$ as a signal on the utility gap $U - U^*$ and of the probability \mathcal{P} as a signal on the chances of success of an application, and that both scores have to be taken into account to design optimal RS. In addition, consistent with intuition, in the last column the fit between job seekers' parameters and vacancies in terms of occupation, reservation wages, skills, diplomas, geographic mobility, significantly predicts that an application is more likely. The only unexpected result here is that the fit in terms of experience in the occupation seems to enter negatively into the application decision.

The introduction of fixed effects forces to restrict the sample to so-called "movers" for whom at least two clicks are observed, including at least one application and one non-application. Thus, to track the changes due to the different specification and the different sample, appendix table C.5 compares the results of the model with fixed effects on movers (column 3) to the ones with uniform weights (column 2) as well as the results without fixed effects on all the population (column 1). Despite the sharp reduction in the number of observations used between these columns the results are close. The table shows the robustness of the result for $1/\mathcal{P}(i, j)$. The estimated coefficients are all negative, as expected, and very close to each other.

4 Empirical evaluation of the different RSs

In this section, we evaluate the alternative RSs based on the performance indicators which are motivated by our search model of Section 3. For each individual i , we consider the set of k_i vacancies, denoted $\mathcal{V}_i(B)$, on which he or she has clicked. These individual sets of vacancies constitute what we call the "benchmark" against which we will compare the different RSs. First, we consider the two RSs presented in section 1, based on $\mathcal{U}(i, j)$ and $\mathcal{P}(i, j)$ respectively. We also explore additional RSs (that we detail below), including the optimal RS described in proposition 3.3, an application-based RS using the score $\Delta(i, j)$, and a pseudo-optimal RS which is a practical implementation of the optimal one. For each of these RSs, generically denoted by S , we select the k_i best vacancies, forming a set of $\mathcal{V}_i(S)$ vacancies.

First, utilizing the FE-logit estimate in column (2) of Table 1.4, we evaluate the quantity that is written here as $\widehat{\Delta(i, j)}/\sigma$, to predict applications:

$$\widehat{\Delta(i, j)}/\sigma = \widehat{\alpha}\mathcal{U}(i, j) - \widehat{\beta}/\mathcal{P}(i, j) + \widehat{\gamma}_i. \quad (1.20)$$

The calculation of averages over the population of functions of $\widehat{\Delta(i, j)}/\sigma$, as shown by the previous expression, requires an estimate $\widehat{\gamma}_i$ of the individual effect in the model (1.19) [see, e.g., Fernández-Val and Weidner, 2018, for a survey].^{26,27}

Second, we compute the function used to value the vacancies, central to the model, which we write here simply as $\widehat{\Gamma}^m(i, j)$ and the associated score, defining the optimal RS $\widehat{S}^*(i, j)$:

$$\widehat{S}^*(i, j) \equiv \widehat{\Gamma}^m(i, j) = \widehat{\sigma} \mathcal{P}(i, j) \log \left(1 + e^{\widehat{\Delta(i, j)}/\sigma} \right), \quad (1.21)$$

where $\widehat{\sigma} = 1/\widehat{\alpha}$.

As stressed before, we consider several RSs: $S \in \{\mathcal{U}, \mathcal{P}, \mathcal{S}^*, \Delta\}$, to which we add a *Pseudo-optimal* one. This pseudo-optimal RS is based on the score similar as the one described in equation (1.21), but using for the Δ the estimates without fixed effects appearing in column (1) of Table 1.4.

4.1 Evaluation metrics

Expected value The procedure for deriving the application-based and optimal recommendation scores as well as the function Γ to value the recommendation sets is as follows.

In order to evaluate these different RSs, we are ultimately interested in what we call the “*expected value*” of the RS:

$$G_i^m(S) = \frac{1}{k_i} \sum_{j \in \mathcal{V}(S)_i} \widehat{\Gamma}^m(i, j)$$

that we compute for each individual for all recommended sets $\mathcal{V}(S)_i$, and on the benchmark set $\mathcal{V}(B)_i$ of vacancies, keeping the search effort in terms of applications k_i con-

²⁶As it is systematically the case with such panel logit models, this individual effect cannot be identified for all job seekers. It is necessary to restrict the analysis to job seekers with at least two clicks, one of which resulted in a job application and the other in which it did not. This introduces a selection of individuals on which we can evaluate the performance of the various RSs. Descriptive statistics on these two full and selected samples are displayed on Table C.4 and show that this selection is limited, except that on the selected subsample the number of observed applications is much larger (median of 12 instead of 3 for the whole sample).

²⁷The quantities of interest reported in Table 1.5 below are averages of moments of these fixed effects γ_i , hence we use a bias correction approach for the estimation [see Hahn and Newey, 2004, Stammann et al., 2016].

stant^{28,29}.

Focusing on our primary outcome variable, the expected value G_i^m , and drawing from the evaluation literature, we identify six potential outcomes corresponding to the benchmark and the five recommender systems under consideration. A key advantage of our framework is the ability to measure all of these potential outcomes for each individual. We report key features of the distribution of these outcomes, as well as the differences between them. This analysis can be conducted for the full sample, where we present the *Average Treatment Effect (ATE)* or the standard deviation of the treatment effects. However, it can also be applied to specific subsamples. Since we are able to compute all potential outcomes and treatment effects for each individual, we can define these subsamples based on either the potential outcomes or the treatment effects.

Overlaps Another important aspect of comparing recommender systems is measuring the overlap rate between the recommendation sets generated by different RSs. For two systems, S_A and S_B , each individual i receives recommendation sets $\mathcal{V}(S_A)_i$ and $\mathcal{V}(S_B)_i$, both of which contain k_i vacancies by design. The overlap rate of the two sets A and B for individual i is calculated as follows:

$$\mathcal{O}_i(A, B) = \frac{\#\{\mathcal{V}(S_A)_i \cap \mathcal{V}(S_B)_i\}}{k_i} \quad (1.22)$$

Similarly, another important dimension of a recommendation system is its impact on the search scope. Each job seeker has a primary occupation, denoted as Occ_i , and there is a corresponding set of available job vacancies in this occupation, represented by \mathcal{A}_i . We can calculate the proportion of vacancies in a set $\mathcal{V}(S)_i$, whether recommended or from the benchmark, that fall within this occupation:

$$\mathcal{O}_i(\mathcal{A}, S) = \frac{\#\{\mathcal{A}_i \cap \mathcal{V}(S)_i\}}{k_i} \quad (1.23)$$

²⁸We talk about “*expected value*”, but this is a shorthand term. It is just one component of the value function as seen from equation 1.16. The component we consider is the empirical analogue of $E(\Gamma^m(P, U) | F_S(S(P, U)) > 1 - s)$, but it enters the value function with a multiplicative $\frac{\alpha_1}{r+q}$ and an additive factor $u(b)$.

²⁹We are also interested in the comparison of the individual hiring rate:

$$\theta_i(S) = \frac{1}{k_i} \sum_{j \in \mathcal{V}(S)_i} \mathcal{P}(i, j) F(\widehat{\Delta(i, j)} / \sigma)$$

on which we report in the appendix.

4.2 Comparative performance of the RSs

Expected value We begin by describing the benchmark, i.e., the distribution of the expected value and the expected hiring rate for job seekers based on the vacancies they clicked on. The average expected value of the vacancies clicked on by job seekers is 0.0105. While this value is not directly interpretable, it serves as a useful reference for comparison.³⁰ The table highlights the heterogeneity in job market prospects among job seekers. The standard deviation of the expected value distribution is 0.0122, which exceeds the mean. We also calculate the average expected value for the bottom 10% and top 10% of individuals in the distribution. For the bottom 10%, the average expected value is 0.0006, which is 16 times lower than the mean. Conversely, for the top 10%, the average expected value is 0.0399, almost four times higher than the mean. This high level of heterogeneity is linked to the variability in hiring rates, which, unlike the expected value, is directly interpretable (see Appendix Table C.6). The standard deviation of the hiring rate is 0.0082, with a mean of 0.0088.³¹ There are significant differences in the average hiring rate across individuals' benchmark sets: the average hiring rate for the bottom 10% of the expected value distribution is 0.0007, which is 12.5 times lower than the average, while for the top 10%, it is 0.0276, or 3 times higher than the average.

Table 1.6 documents the treatment effects of considering one RS rather than the benchmark, or one RS rather than the optimal RS. Thanks to the developed model and the estimated behaviors, we can calculate each of the five potential outcomes for each individual. We are therefore in a situation where, for each individual, we can calculate the treatment effects of one RS in relation to the benchmark, and of one RS in relation to another. Given this particular situation, it is possible to calculate not only an average treatment effect, but also the distributional characteristics of the individual treatment effect. The table gives the average impact on the total population (column (1)) and on individuals at the top and bottom of the benchmark distribution (columns (4) and (5)).³² It also gives the standard deviation of the treatment effect distribution (column (6)), as well as the mean of the recommenders' effects on the top and bottom 10% of the distribution of the effect of switching from the benchmark to the the optimal RS (columns (2) and (3)).

The first result in the table is that the optimal RS substantially improves the situation

³⁰If both the utility and error term ε are multiplied by the same factor λ , the ratio $\Delta(u, p)/\sigma$ remains unaffected because both Δ and σ are multiplied by λ , but the expected value is scaled by λ through its effect on σ .

³¹Although this hiring rate appears quite low, it is important to note that it accounts for both the probability of application and the probability of hiring conditional on an application. The probability of hiring conditional on an application is discussed in Section 2.3 for the best vacancy recommended by the \mathcal{U} and \mathcal{P} RSs. For the latter, the median is approximately 0.06.

³²This information is already present in Table 1.5, which shows the averages for each of the RSs separately, it is more directly accessible here.

Table 1.5: Average expected value of the different RSs

	Average over:			
	All (1)	Bottom 10% (2)	Top 10% (3)	Std. dev. (4)
Expected value: G^m (x100)				
Benchmark	1.05	0.06	3.99	1.22
\mathcal{U}	0.52	0.03	1.87	0.78
\mathcal{P}	1.90	0.13	6.29	2.18
Δ	1.89	0.14	6.02	2.10
Optimal	2.06	0.15	6.69	2.29
Pseudo	2.06	0.15	6.64	2.27

(1) is the average over the full sample.

(2) (resp. (3)) is the average over these in the bottom (resp. top) 10% of the distribution of the expected value.

(4) is the standard deviation over the full sample.

of job seekers on average. The average treatment effect (see table 1.6) is 0.0102, which represent 77% of the standard deviation of the value function in the benchmark 0.0122 (see table 1.5). On average, the expected value is almost doubled: from 0.0105 to 0.0206.³³ This improvement is also seen for job seekers at the lower end of the benchmark expected value distribution: for them, too, the expected value doubles, from 0.0006 to 0.0015. Even though this is a doubling, the improvement in these job seekers' prospects remains modest. We also observe that at the top of the benchmark distribution, the average treatment effect is 0.0270, i.e. 30 times higher than at the bottom. We conclude that, while the optimal RS benefits everyone, it is not likely to reduce inequalities on the labour market. On the contrary, it seems to benefit mainly those in the best position. Still with a view to documenting the heterogeneity of the impact, we observe that the standard deviation of the treatment effect (table 1.6 column (6)) is 0.0134. It is therefore greater than the average treatment effect. We also observe that the average treatment effect of the optimal recommendations for those for whom it is in the bottom 10% of the treatment effect distribution is very low (0.0006), i.e. 17 times lower than the average effect. On the other hand, it is very large at the top of the treatment effect distribution (0.0411), i.e. more than 4 times the average effect and 70 times the effect at the bottom of the distribution.

Interestingly the performance of the Pseudo optimal RS is very close. The last line of Table 1.6 shows that the average and standard deviation of the treatment effect associated

³³A similar progression can be observed for the hiring rate, which rises from 0.0088 to 0.00160 on average. See appendix Table C.6 which presents impacts on hiring.

with switching from the Optimal to the Pseudo-optimal RS are very small (respectively 0.0001 and 0.0004).

The table also shows that the RSs based on recruitment chances $\mathcal{P}(i, j)$ or the one based on applications $\Delta(i, j)$ have very similar performances. In addition, their performance are only slightly below than those of the optimal RS. The average expected values achieved by these two RSs are respectively 0.0190 and 0.0189 (table 1.5). This is 9% below the level reached with the optimal RS (2.06). At the lower end of the benchmark distribution, the three RSs perform very closely together. There are differences between the three recommendation systems, however, but they remain limited. These differences are at the upper end of the benchmark distribution. The optimal RS (0.0669) does slightly better than that based on \mathcal{P} (0.0629), which in turn does slightly better than that based on Δ (0.602). Table 1.6 confirms the similarity of performance between recommendations based on \mathcal{P} , Δ and the optimal RS, whose treatment effects (compared to Benchmark) are respectively 0.085 and 0.084.

Another salient result is the poor performance of the RS based on \mathcal{U} . Table 1.6 clearly shows that not only the average treatment effect relative to the benchmark is negative, but that this worsening of labor market prospects by this RS is observed in all subpopulations examined: the top and bottom 10% of the benchmark, but also the top and bottom 10% of the distribution of the effect of switching from the benchmark to the optimal RS. Similar results can be observed for the hiring rate: the \mathcal{U} RS systematically worsens the chances of finding a new job. If we compare the RSs based on $\mathcal{P}(i, j)$ with those based on $\mathcal{U}(i, j)$, one of our initial concerns, we see that the RS based on $\mathcal{P}(i, j)$ clearly dominates.

The poor performance of the RS based on \mathcal{U} does not mean that the utility perceived by job seekers is not a relevant dimension for making recommendations. Rather, it highlights the relatively poor quality of the \mathcal{U} utility indicator. Indeed, the standard deviation of the $\mathcal{U}(i, j)$ score on vacancies clicked by job seekers is 0.225 and 0.101 when we calculate the “within” standard deviation. However, our estimate of the σ parameter is $1/1.10 = 0.91$, i.e. a standard deviation for ε of 0.91 (see proposition 3.2). This is a large value compared to the within standard deviation. If we had a better measure of the utility of a job, taking better account of the relevant characteristics of a job, the standard deviation of ε would be lower and the optimal score would tend towards $\mathcal{UP} - \beta$. So if \mathcal{U} does not seem to play an important role in our results, it might be because the measure of job quality is noisy.

Overlaps Table 1.7 presents the results regarding the overlap rate of different recommendation sets. The overall finding is that these recommendation sets are generally very different, with the notable exception of the optimal RS and the pseudo-optimal RS, which have an average overlap rate of 0.98. This high overlap rate confirms the similarity be-

Table 1.6: Comparison of the expected value of the different RSs

	Average effect over:					Std. Dev. (6)
	All (1)	Optimal - benchmark effect		Benchmark		
		Bottom 10% (2)	Top 10% (3)	Bottom 10% (4)	Top 10% (5)	
Expected value: G^m (x100)						
Optimal - Benchmark	1.02	0.06	4.11	0.09	2.70	1.34
\mathcal{P} - Benchmark	0.85	0.04	3.71	0.07	2.30	1.25
Δ - Benchmark	0.84	0.05	3.46	0.08	2.03	1.20
\mathcal{U} - Benchmark	-0.52	-0.05	-1.33	-0.03	-2.12	0.87
Optimal - \mathcal{P}	0.17	0.02	0.40	0.02	0.40	0.20
Optimal - Δ	0.17	0.01	0.64	0.01	0.66	0.31
Optimal - \mathcal{U}	1.54	0.11	5.44	0.12	4.81	1.79
Optimal - Pseudo	0.01	0.00	0.05	0.00	0.05	0.04

(1) is the average effect over the full sample.

(2) (resp. (3)) is the average effect over those in the bottom (resp. top) 10% of the distribution of the effect of switching from the benchmark to the optimal RS on expected value.

(4) (resp. (5)) is the average effect over those in the bottom (resp. top) 10% of the expected value distribution of the benchmark.

(6) is the standard deviation over the full sample.

tween the two recommendation systems in terms of their performance, as measured by Γ . It also confirms that it is possible to consider the recommendation set obtained from the estimation without accounting for fixed effects, which has important practical implications. However, for the other RSs, we observe very low overlap rates. For example, the overlap rate between the benchmark and the vacancies from any of the five RSs considered does not exceed 14%. Similarly, the overlap between the recommendations from the \mathcal{P} -based system and those from the optimal RS is 57%, despite the similar performance of the two RSs. The same applies to the RS based on applications (Δ), where the overlap with the optimal RS is only 67%, and with the RS based on \mathcal{P} , it is just 30%, again despite their similar performance. The recommendation system based on \mathcal{U} stands out as the most distinct from the others. The highest overlap with the application-based system (Δ) is only 14%, while it is merely a few percent for the other systems.

Appendix Table C.8 presents the results regarding the extent of the search area beyond the job corresponding to the primary occupation sought. For the benchmark, we observe that only 30% of the vacancies clicked on by job seekers actually correspond to their primary occupation. This suggests that job seekers spontaneously diversify their search substantially to include other occupations. A striking and unexpected result is that all RSs tend to refocus the search activity on the primary occupation. Logically, at

Table 1.7: Similarity rates on recommended vacancies

	Base	Optimal	Pseudo	\mathcal{P}	\mathcal{U}	Δ
Base	1.00	0.14	0.14	0.12	0.01	0.11
Optimal		1.00	0.98	0.57	0.07	0.67
Pseudo			1.00	0.56	0.07	0.69
\mathcal{P}				1.00	0.02	0.30
\mathcal{U}					1.00	0.14
Δ						1.00

Note: these are average rates of overlap on recommended vacancies between the system in the column and the one on the row, taken over the job seekers in our sample and weighted by the number of individual clicks (17,865 in total, with a median of 12 per job seeker).

one end of the spectrum are the recommendations based on \mathcal{U} , which are 65% concentrated on the main sector of activity. Conversely, the RS based on hiring chances (\mathcal{P}), as expected, diversifies recommendations more towards other sectors, with an overlap rate of 32% (interestingly, this rate is lower than that of the job seekers' own search behavior). The optimal recommendation system and the one based on applications (Δ) fall in between, and it is worth noting that, in this regard, the optimal recommendation system is closer to the system based on \mathcal{U} than to \mathcal{P} .

Conclusion: implications for designing RSs in practice

Recommendation algorithms are becoming an integral part of PES, with many institutions planning to adopt them in the near future [see Broecke, 2023]. However, there is no one-size-fits-all approach to designing these systems [Freire and de Castro, 2021], as different methodologies yield varying outcomes depending on the specific objectives and constraints of the recommendation system. In this paper, we explore the economic aspects of job RSs, focusing on how they can effectively support job seekers in navigating the labor market.

Our analysis shows that these algorithms effectively target latent factors, which, when properly combined, can generate optimal recommendations for job seekers. By integrating knowledge of hiring probabilities with other relevant factors, these systems significantly improve the job search process by helping seekers identify opportunities where they are likely to apply and be hired. An important finding is that the optimal RS emerges from a combination of two key components: the utility (\mathcal{U}) job seekers derive from various job attributes and the probability (\mathcal{P}) of a successful application. However, while ML

tasks can predict \mathcal{P} with reasonable accuracy, predicting utility (\mathcal{U}) is more challenging due to noisy information, such as reservation wage and mobility preferences. However, one can notice from equation (1.13) that the optimal RS simply requests to identify the score Δ that could be directly identified thanks to a ML task dedicated to predict applications and the \mathcal{P} score. Combining the two latter quantities to get the score forming the optimal RS is then quite straightforward. This approach demonstrates the potential for ML to learn deep, hard-to-measure quantities, which, when integrated into an economic framework, produce an optimal recommendation score. We identify two potential refinements of our baseline analysis: adjusting the reservation utility in response to recommendations and relaxing the rationality assumption. Yet, we find that combining identified quantities from our baseline approach remains paramount over these refinements.

Our analysis points to promising directions for enhancing recommendation systems by better capturing job seekers' utility for job attributes. This can be achieved through the use of hypothetical scenarios, as explored in works like Mas and Pallais [2017] and Wiswall and Zafar [2018], Banerjee et al. [2022], Banerjee and Chiplunkar [2024]. Additionally, leveraging repeated interactions, as seen in bandit literature for matching markets [see, e.g., Jagadeesan et al., 2021], could further refine these systems, ensuring that recommendations are aligned with objective expected utility rather than merely replicating past behaviors. Moving forward, it will be crucial to validate our theoretical findings in real-world settings. An important question remains: can job seekers discern the values embedded in different recommendation sets and adjust their search strategies accordingly? The goal is to conduct a detailed analysis of the data from the March 2022 beta test, which will allow us to fully address these questions. The results presented in this chapter are preliminary and serve as an initial exploration of the dataset.

A key area of future research is to examine how these recommendation systems impact real-world labor market outcomes, such as job application behavior, successful placements, and employment quality. This will require larger sample sizes and continuous participation from job seekers interacting with the recommendation system. A comprehensive, large-scale study with long-term exposure to recommendation streams is essential, incorporating experimental designs that account for potential spillover effects. Further research should also compare in the field the impacts of ML-, preference-based and hybrid RSs across different job seeker groups to identify which segments benefit most from each type of algorithm. Additionally, refining the \mathcal{P} -REC system using direct feedback from job seekers is a promising path to improve its effectiveness. It is equally important to address fairness and inequality issues that arise from RSs in labor markets, as well as concerns related to congestion [Zhang and Kuhn, 2022, Bied et al., 2023a, Altmann et al., 2022, Bied et al., 2021b, Le Barbanchon et al., 2023, Behaghel et al., 2024, Lehmann et al.,

2023]. Lastly, extending recommendation systems to benefit firms presents another valuable direction for future research [Horton, 2017, Algan et al., 2020].

References

- Yann Algan, Bruno Crépon, and Dylan Glover. Are active labor market policies directed at firms effective? evidence from a randomized evaluation with local employment agencies. J-PAL working paper, 2020.
- Steffen Altmann, Armin Falk, Simon Jäger, and Florian Zimmermann. Learning about job search: A field experiment with job seekers in germany. Journal of Public Economics, 164:33–49, 2018.
- Steffen Altmann, Anita M Glenney, Robert Mahlstedt, and Alexander Sebald. The direct and indirect effects of online job search advice. 2022.
- David H Autor. Wiring the labor market. The Journal of Economic Perspectives, 15(1): 25–40, 2001.
- Linda Babcock, William J Congdon, Lawrence F Katz, and Sendhil Mullainathan. Notes on behavioral economics and labor market policy. IZA Journal of Labor Policy, 1:1–14, 2012.
- Mirjam Bächli, Hélène Benghalem, Doriana Tinello, Damaris Aschwanden, Sascha Zuber, Matthias Kliegel, Michele Pellizzari, and Rafael Lalive. Ranking occupations by their proximity to workers’ profiles. Swiss Journal of Economics and Statistics, 160(1):1–17, 2024.
- A. Banerjee, B. Crépon, E. Perennes, and C. Welter-Médée. Using stated individual preferences for job ads’ attributes to design a better matching algorithm between job ads and job-seekers. AEA RCT Registry, march 2022. doi: 10.1257/rct.8719.
- Abhijit V Banerjee and Gaurav Chiplunkar. How important are matching frictions in the labor market? experimental & non-experimental evidence from a large indian firm. Journal of Development Economics, page 103330, 2024.
- Luc Behaghel, Sofia Dromundo, Marc Gurgand, Yagan Hazard, and Thomas Zuber. The potential of recommender systems for directing job search: A large-scale experiment. 2024.
- Michèle Belot, Philipp Kircher, and Paul Muller. Do the long-term unemployed benefit from automated occupational advice during online job search? 2022.
- Michèle Belot, Philipp Kircher, and Paul Muller. Providing advice to jobseekers at low cost: An experimental study on online advice. The review of economic studies, 86(4): 1411–1447, 2019.

- Guillaume Bied, Philippe Caillou, Bruno Crépon, Christophe Gaillac, Victor Alfonso Naya, Elia Pérennes, and Michèle Sebag. Designing labor market recommender systems: the importance of job seeker preferences and competition. In 4. IDSC of IZA Workshop: Matching Workers and Jobs Online-New Developments and Opportunities for Social Science and Practice, 2021a.
- Guillaume Bied, Elia Pérennes, Victor Alfonso Naya, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michele Sebag. Congestion-avoiding job recommendation with optimal transport. In FEAST workshop ECML-PKDD 2021, 2021b.
- Guillaume Bied, Christophe Gaillac, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Solal Nathan, and Michèle Sebag. Fairness in job recommendations: Estimating, explaining, and reducing gender gaps. forthcoming in the proceedings of ECAI-workshop AEQUITAS, 2023a.
- Guillaume Bied, Solal Nathan, Elia Pérennes, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michèle Sebag. Toward Job Recommendation for All. In Thirty-Second International Joint Conference on Artificial Intelligence {IJCAI-23}, pages 5906–5914, Macau, China, August 2023b. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2023/655. URL <https://hal.science/hal-04245528>.
- Stijn Broecke. Oecd social, employment and migration working papers: Artificial intelligence and labour market matching. 2023.
- Kuan-Ming Chen, Yu-Wei Hsieh, and Ming-Jen Lin. Reducing recommendation inequality via two-sided matching: a field experiment of online dating. International Economic Review, 2023.
- Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018.
- Michael Cooper and Peter Kuhn. Behavioral job search. Handbook of Labor, Human Resources and Population Economics, pages 1–22, 2020.
- Corné De Ruijt and Sandjai Bhulai. Job recommender systems: A review. arXiv preprint arXiv:2111.13576, 2021.
- Iván Fernández-Val and Martin Weidner. Fixed effects estimation of large-t panel data models. Annual Review of Economics, 10(1):109–138, 2018.

- Erica Field, Robert Garlick, Nivedhitha Subramanian, and Kate Vyborny. Why don't jobseekers search more? barriers and returns to search on a job matching platform. working paper, 2023.
- Mauricio Noris Freire and Leandro Nunes de Castro. e-recruitment recommender systems: a systematic review. Knowledge and Information Systems, 63:1–20, 2021.
- Dylan Glover. Job search and intermediation under discrimination: Evidence from terrorist attacks in france. Chaire Securisation des Parcours Professionels Working Paper,(2019-02), 164, 2019.
- Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016. <http://www.deeplearningbook.org>.
- Francisco Gutiérrez, Sven Charleer, Robin De Croon, Nyi Nyi Htun, Gerd Goetschalckx, and Katrien Verbert. Explaining and exploring job recommendations: a user-driven approach for interacting with knowledge-based job recommender systems. In Proceedings of the 13th ACM Conference on Recommender Systems, pages 60–68, 2019.
- Jinyong Hahn and Whitney Newey. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica, 72(4):1295–1319, 2004.
- Gunter J Hitsch, Ali Hortaçsu, and Dan Ariely. Matching and sorting in online dating. American Economic Review, 100(1):130–63, 2010.
- John J Horton. The effects of algorithmic labor market recommendations: Evidence from a field experiment. Journal of Labor Economics, 35(2):345–385, 2017.
- Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. Advances in Neural Information Processing Systems, 34:3323–3335, 2021.
- Maximilian Kasy. The political economy of ai regulation: Towards democratic control of the means of prediction. Working paper, 2023.
- Erin M Kelley, Christopher Ksoll, and Jeremy Magruder. How do digital platforms affect employment and job search? evidence from india. Journal of Development Economics, 166:103176, 2024.
- Philipp Kircher. Job search in the 21st century. Journal of the European Economic Association, 20(6):2317–2352, 2022.

- Philipp AT Kircher. Search design and online job search—new avenues for applied and experimental research. Labour economics, 64:101820, 2020.
- Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences, 118(22), 2021.
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. arXiv preprint arXiv:2202.11776, 2022.
- Peter Kuhn and Hani Mansour. Is internet job search still ineffective? The Economic Journal, 124(581):1213–1233, 2014.
- Thomas Le Barbanchon, Roland Rathelot, and Alexandra Roulet. Gender differences in job search: Trading off commute against wage. The Quarterly Journal of Economics, 136(1):381–426, 2021.
- Thomas Le Barbanchon, Lena Hensvik, and Roland Rathelot. How can ai improve search and matching? evidence from 59 million personalized job recommendations. Technical report, Working Paper, 2023.
- Tobias Lehmann, Camille Terrier, and Rafael Lalive. Improving matching efficiency in two-sided markets: A mutual popularity ranking approach. working paper, 2023.
- Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to match via inverse optimal transport. J. Mach. Learn. Res., 20(80):1–37, 2019.
- Shichuan Ma, Haiyan Luo, Jianjie Ma, Ziyang Liu, Yu Sun, Xingang Huang, Fengdan Wan, Veeresh Beeram, Henry Oh, Santosh Raghuram Kumar, et al. Jobs filter to improve the job seeker experience at indeed. com. 2022.
- Ioana Marinescu and Daphné Skandalis. Unemployment insurance and job search behavior. The Quarterly Journal of Economics, 136(2):887–931, 2021.
- Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. American Economic Review, 107(12):3722–59, 2017.
- Yoosof Mashayekhi, Nan Li, Bo Kang, Jeffrey Lijffijt, and Tijl De Bie. A challenge-based survey of e-recruitment recommendation systems. arXiv preprint arXiv:2209.05112, 2022.
- Yoosof Mashayekhi, Bo Kang, Jeffrey Lijffijt, and Tijl De Bie. Recon: Reducing congestion in job recommendation using optimal transport. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 696–701, 2023.

- Andreas I Mueller and Johannes Spinnewijn. Expectations data, labor market, and job search. Handbook of Economic Expectations, pages 677–713, 2023.
- Andreas I Mueller, Johannes Spinnewijn, and Giorgio Topa. Job seekers’ perceptions and employment prospects: Heterogeneity, duration dependence, and bias. American Economic Review, 111(1):324–363, 2021.
- Jun Shi, Chengming Jiang, Aman Gupta, Mingzhou Zhou, Yunbo Ouyang, Qiang Charles Xiao, Qingquan Song, Yi Wu, Haichao Wei, and Huiji Gao. Generalized deep mixed models. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 3869–3877, 2022.
- Amrei Stammann, Florian Heiss, and Daniel McFadden. Estimating fixed effects logit models with large panel data. 2016.
- Gerhard Tutz, Matthias Schmid, et al. Modeling discrete time-to-event data. Springer, 2016.
- Gerard J Van den Berg and Bas Van der Klaauw. Counseling and monitoring of unemployed workers: Theory and evidence from a controlled social experiment. International economic review, 47(3):895–936, 2006.
- Bas van der Klaauw and Heike Vethaak. Empirical evaluation of broader job search requirements for unemployed workers. Technical report, Tinbergen Institute Discussion Paper, 2022.
- Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. Content-based neighbor models for cold start in recommender systems. In Proceedings of the Recommender Systems Challenge 2017 - RecSys Challenge 17. ACM Press, 2017. doi: 10.1145/3124791.3124792.
- Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res., 10(2), 2009.
- Halbert White. Maximum likelihood estimation of misspecified models. Econometrica: Journal of the econometric society, pages 1–25, 1982.
- Matthew Wiswall and Basit Zafar. Preference for the workplace, investment in human capital, and gender. The Quarterly Journal of Economics, 133(1):457–507, 2018.
- Shuo Zhang and Peter Kuhn. Understanding algorithmic bias in job recommender systems: An audit study approach. 2022.

Jing Zhao, Jingya Wang, Madhav Sigdel, Bopeng Zhang, Phuong Hoang, Mengshu Liu, and Mohammed Korayem. Embedding-based recommender system for job to candidate matching on scale. arXiv preprint arXiv:2107.00221, 2021.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. Advances in Neural Information Processing Systems, 33:15763–15773, 2020.

Appendices

A Field experiment

This appendix provides details on algorithms used for the March 2022 field experiment, as well as supplementary material for its analysis.

A.1 Algorithms

The \mathcal{U} -REC algorithm The \mathcal{U} -REC algorithm takes the form:

$$s_{ij} = f_{ij} \times \left(\sum_{k \in \text{Criteria}} w_k s_{ijk} \right)$$

where:

- $f_{ij} \in \{0, 1\}$ is a filter, taking value 1 if the score assigned to geography is different from zero
- The criteria s_{ijk} and their weights w_k are:
 1. Skills: weight 1000
 2. Job type: weight 1500 (500 at the ROME level and 1000 at a finer level of granularity)
 3. Experience: weight 100
 4. Geographic mobility³⁴, weight 100
 5. Contract type, weight 10
 6. Weekly working hours, weight 100
 7. Education, weight 100
 8. Languages, weight 100
 9. Driver's licence, weight 100
 10. Wage, weight 200

Scores take value between 0 (complete mismatch) and 1 (perfect fit), with intermediate values determined by expert-provided matrices and discontinuities. The final ranking is a lexicographic sort by decreasing s_{ij} , increasing geographic distance (to one's zip code of residence), and decreasing job ad creation date.

³⁴This score is not necessarily based on distance between the ad and the job seekers' place of residence. It takes into account the kind of mobility declared acceptable by the job seeker. Job seekers can declare a zip code and a commuting radius (in which case the score is a decreasing function of distance, taking value 0 after a threshold), but also, less often, a country, region or *département* (in which case the score takes binary values).

The MIX algorithm We first attribute “stars” to job ads with respect to both \mathcal{P} -REC.0 and \mathcal{U} -REC to construct a consideration set of job ads that have a high ranking for one of these algorithms, or good rankings for both. Stars with respect to an algorithm are determined in the following fashion: 4 if the ad’s rank is below 10; 3 if the ad’s rank is below 25; 2 if the ad’s rank is below 50; 1 if the ad’s rank is below 100; 0 otherwise.

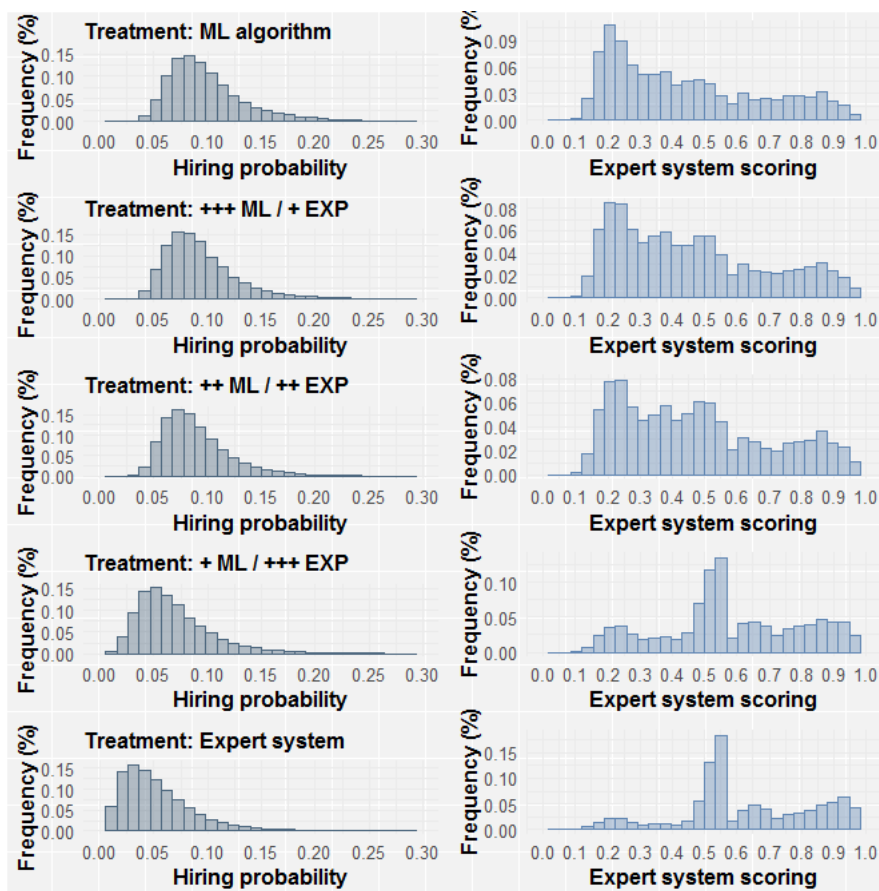
MIX only takes into consideration job ads for which the sum of \mathcal{P} -REC.0 and \mathcal{U} -REC “stars” are greater or equal to 3. This consideration set takes a size between 25 (the top-25s of \mathcal{P} -REC.0 and \mathcal{P} -REC.1 are the same) and 100 (disjoint top-25s, and the job ads ranked 25-50 by an algorithm are among the other’s top 50-100).

From this consideration set, MIX aims to generate 15 recommendations per job seeker³⁵. MIX- p ($p \in \{1/4, 1/2, 3/4\}$) takes the $\max(15, p \times \text{size}(\text{consideration set}))$ first ones according to \mathcal{P} -REC.0, and reorders them by the \mathcal{U} -REC score.

³⁵In order to present 10 job ads to each job seeker in the experiment. A larger amount of ads nevertheless has to be ranked in order to anticipate a mismatch between job ads available at recommendation time and those actually online at the time of sending the survey (in order to make sure recommended job ads are actually online).

A.2 Differences in recommendations policies between algorithms

Figure A.1: Scores distributions



Note: These figures display the distribution of $\mathcal{P}(i, j)$ and $\mathcal{U}(i, j)$ scores of recommended job ads among participants to the field experiment, for all five algorithms. “ML algorithm” refers to \mathcal{P} -REC, “+++ ML/+EXP” refers to MIX^{-1/4}, “++ML/++EXP” refers to MIX^{-1/2}, “+ML/+++EXP” refers to MIX^{-3/4} and “Expert system” refers to \mathcal{U} -REC.

A.3 Survey design

Surveyed population The eligible population are job seekers registered at France Travail in the Auvergne-Rhône-Alpes region, of administrative category A (i.e. available for a job and looking for one), aged over 18 years old, and having given the PES the permission to contact them by email. Randomization was stratified by desired job type (14 modalities), the kind of accompaniment delivered by the institution (3 modalities describing the job seeker's degree of autonomy), and geographic location (level of a French *département*, 12 modalities).

Survey protocol The experiment was conducted in March 2022. Job seekers are sent an email inviting them to complete an online survey (screenshots below). A link provided in the email directs them to the survey's cover page. The cover page provides them information on the survey's goals, as well as assurance that the information collected will be used for research purposes and have no impact on their treatment by France Travail.




If they accept those terms, job seekers are first shown two job ads (their assigned algorithm's top-2). Job ads are characterized by the firm, working conditions, wage, workplace (and distance), experience, educational requirements, driver's license requirements, and an overview of the job's and firm's textual description. Job seekers are asked to rate the two job ads (out of ten, on a continuous slider) in terms of i) global relevance, ii) their perception of their chances of being recruited, and iii) fit to their job search criteria. They may also optionally provide comments in natural language.

After rating the two job ads (which is mandatory to proceed in the survey), job seekers visualize a final page displaying ten job ads (their assigned algorithm's top-10, including the two previously seen ones). Job seekers do not have to rate ads on this page. They may click on the ads to view them on France Travail's website (which provides further details on the ads, and allows job seekers to apply if they wish to). Clicks of job seekers on the ads are recorded.

Overview of the survey





Figure A.2: Landing page


Progression 100%


Contribuez à faire évoluer l'offre de service de Pôle emploi en participant à ce test !

Ce rapide questionnaire vous donne l'opportunité de tester un **nouveau service de recommandation d'offres d'emploi** et de nous **donner votre avis** sur les recommandations qui vous sont faites.

-  **Ce test ne vous prendra que 5 minutes !**
Vous allez visualiser des recommandations d'offres d'emploi (en ligne sur pole-emploi.fr) et vous pourrez donner votre avis sur chacune d'elles.
-  **Accédez à des recommandations d'offres personnalisées.**
Les recommandations d'offres d'emploi qui vous sont proposées dans ce test ont été calculées pour vous, en fonction de votre profil.
-  **Ce test est anonyme.**
Vos réponses resteront anonymes et seront analysées à des fins statistiques par des personnes habilitées et soumises au devoir de confidentialité.
-  **Vos réponses ne seront pas transmises à votre conseiller.**
Vos réponses seront analysées uniquement pour cette enquête. Elles n'entraîneront aucune modification de votre dossier Pôle emploi.



Souhaitez-vous participer à ce test ?

 **Commencer le test et accéder à mes recommandations**

En cliquant sur **Commencer**, vous acceptez que les réponses que vous donnerez soient récoltées et analysées par l'équipe de recherche rattachée au CREST (Centre de Recherche en Economie et Statistiques) et au LISN (Laboratoire Interdisciplinaire des Sciences du Numérique) dans le cadre de leur partenariat avec Pôle emploi.

Quelques précisions sur la collecte de données personnelles:

Vos données seront traitées avec le logiciel d'enquête Qualtrics, en conformité avec les exigences de la loi applicable et sur la base de l'exécution de la mission d'intérêt public de Pôle emploi. Votre numéro identifiant a été tiré au sort pour cette enquête. Vos informations sont collectées à compter de l'instant où vous cliquez sur le bouton « Commencer ». Les réponses à l'enquête sont reçues par l'équipe de chercheurs du CREST et du LISN. Elles sont anonymisées et analysées uniquement à des fins statistiques par les personnes habilitées et soumises au devoir de confidentialité. Conformément à la loi applicable, vous pouvez exercer vos droits en vous adressant à courriers-oh@pole-emploi.fr. Si vous avez un doute, vous pouvez également répondre au mail que vous avez reçu de notre part pour nous en parler.

Produit par Qualtrics

Figure A.3: First page

The image shows a job offer page for 'Auxiliaire de puériculture' (Nanny) at 'FAMILLES RURALES D ALIXAN'. The page includes logos for 'pôle emploi', 'CREST', and 'LASH'. The job details are as follows:

- Entreprise:** FAMILLES RURALES D ALIXAN (6 à 9 salariés)
- Conditions:** Contrat à durée déterminée (Durée : 01 mois), 30H Horaires normaux
- Salaires:** Horaire : 10.25 euros
- Lieu de travail:** Alixan (à 14.3 kilomètres)
- Expérience:** Débutant accepté
- Formation:** Diplôme : CAP BEP AUXILIAIRE PUERICULTURE OBLIG. (exigé) / Domaine Auxiliaire puériculture
- Permis:** Non renseigné

Description du poste :
Vous intervenez dans un centre multi-accueil auprès des enfants pour les accompagner et dans les gestes de la vie quotidienne. Missions : - prendre soin de l'enfant dans ses activités quotidiennes, - observer l'enfant et noter les évolutions liées à son développement, - assurer l'immédiat de l'enfant, - recueillir et transmettre ces observations, - accueillir, informer, accompagner l'enfant et sa famille, - réaliser des activités d'éveil, de loisirs et d'éducation, - accueillir et accompagner des collègues et des stagiaires, - utiliser les techniques d'entretien des locaux et du matériel son travail au sein d'une équipe. Poste dans le cadre d'un remplacement DIPLOME AUXILIAIRE PUERICULTURE OBLIGATOIRE (...)

Description de l'entreprise :
17 enfants accueillis de 7h30 à 18h30 (3 mois à 3 ans) Équipe de 6 personnes

Feedback Form:

Question 1 : Globalement, quelle note sur 10 donnez-vous à cette offre ?
0 1 2 3 4 5 6 7 8 9 10
Progress bar: 5

Question 2 : À quel point cette offre convient-elle à vos critères de recherche* ?
*Par critères de recherche nous entendons : le métier recherché, le contrat de travail souhaité, le salaire minimum souhaité, le temps de trajet maximum accepté et la durée hebdomadaire souhaitée (temps plein/temps partiel).
0 : Offre ne convient pas du tout à mes critères | 10 : Offre convient parfaitement à mes critères
0 1 2 3 4 5 6 7 8 9 10
Progress bar: 5
Optionnel : détaillez votre note pour chaque critère

Question 3 : Comment estimez-vous vos chances d'être embauché sur cette offre, si vous y postulez ?
0 : Aucune chance | 10 : Chances particulièrement élevées
0 1 2 3 4 5 6 7 8 9 10
Progress bar: 5

Si vous le souhaitez, vous pouvez aussi nous en dire plus ci-dessous :

Buttons: [Retour à la page d'accueil](#) [Voir l'offre suivante](#)

Produit par Qualitrics

Figure A.4: Second page

The screenshot displays a web interface for job listings on Pôle-emploi.fr. At the top, there are logos for Pôle-emploi, CREST, and LISN. A message in red text says: "Merci d'avoir pris le temps de noter ces deux offres ! Vos réponses ont été enregistrées." Below this, it states: "Nous vous proposons ci-dessous 3 offres d'emploi supplémentaires (pas besoin de les noter cette fois !)" and "Vous pouvez aller consulter ces 5 offres sur Pôle-emploi.fr et y postuler si vous le souhaitez en cliquant sur les liens ci-dessous."

The page lists three job offers, each with a title and a list of criteria:

- Offre 1 / 5: Auxiliaire de puériculture**
 - Entreprise: COPAINS-CRINES 6625 (Sans salaire)
 - Conditions: Contrat à durée déterminée (Durée: 01 mois), 35h Horaires normaux
 - Salaires: Horaire: 10,25 euros
 - Lieu de travail: Guilherand Granges (à 3,4 kilomètres)
 - Expérience: Débutant accepté
 - Formation: Diplôme: CAP SEP (obligé) / Domaine: Petite enfance
 - Permis: Non renseigné
- Offre 2 / 5: Auxiliaire de puériculture**
 - Entreprise: FAMILLES RURALES D'ALLIAN (à 9 salariés)
 - Conditions: Contrat à durée déterminée (Durée: 01 mois), 35h Horaires normaux
 - Salaires: Horaire: 10,25 euros
 - Lieu de travail: Allian (à 14,3 kilomètres)
 - Expérience: Débutant accepté
 - Formation: Diplôme: CAP SEP AUXILIAIRE PUERICULTURE OBLIG. (obligé) / Domaine: Auxiliaire puériculture
 - Permis: Non renseigné
- Offre 3 / 5: Educateur / Educatrice de jeunes enfants**
 - Entreprise: BIRELLAE (à 9 salariés)
 - Conditions: CDI, 35h Horaires normaux
 - Salaires: Mensuel: 1800,0 euros
 - Lieu de travail: Solence (à 7,4 kilomètres)

Each offer includes a link to "Voir le détail de l'offre sur Pôle-emploi.fr".

On the right side of the page, there are two more job offers partially visible:

- Offre 4 / 5: Assistant / Assistante accueil petite enfance**
 - Entreprise: CRED-ES EXPANSION DROUVE-ARDECHE (3 à 5 salariés)
 - Conditions: Contrat à durée déterminée (Durée: 01 mois), 35h Horaires normaux
 - Salaires: SMIC
 - Lieu de travail: Montier (à 14,9 kilomètres)
 - Expérience: Débutant accepté
 - Formation: Diplôme: CAP SEP (obligé) / Domaine: Petite enfance
 - Permis: B - Véhicule léger (souhaité)
- Offre 5 / 5: Secrétaire médical / médicale**
 - Entreprise: IMAGERIE MEDICALE ET RADIOTHERAPE (20 à 49 salariés)
 - Conditions: Contrat à durée déterminée (Durée: 08 mois), 35h Horaires normaux
 - Salaires: Mensuel de 1600,0 euros à 1450,0 euros
 - Lieu de travail: Guilherand Granges (à 3,4 kilomètres)
 - Expérience: 2 ans d'expérience minimum
 - Formation: Non renseigné
 - Permis: Non renseigné

At the bottom right, there is a small box with the text: "Si vous le souhaitez, vous pouvez nous donner votre avis sur ces 3 recommandations supplémentaires, sur le service de recommandation en général ou sur ce test:"

Note: Clicking on “Voir le détail de l’offre sur Pôle-emploi.fr” leads to a more thorough description of job ads on France Travail’s website, on which job seekers may also apply.

A.4 Randomization check

Table A.1: Balance check among full sample

	\mathcal{P} -REC	MIX- 1/4	MIX- 1/2	MIX- 3/4	\mathcal{U} -REC	p
Age	38.18	38.09	38.46	38.38	38.47	0.10
Looking for a permanent contract, full time	0.65	0.65	0.65	0.65	0.64	0.66
Looking for a permanent contract, part time	0.11	0.11	0.11	0.11	0.12	0.94
Looking for a temporary contract	0.19	0.19	0.19	0.20	0.19	0.77
Education: High school	0.25	0.26	0.26	0.25	0.26	0.56
Education: Less than high school	0.10	0.10	0.10	0.10	0.10	0.71
Education: Vocational training	0.28	0.27	0.28	0.27	0.27	0.25
Education: College Education	0.37	0.37	0.37	0.37	0.37	0.95
Gender: Woman	0.53	0.52	0.52	0.54	0.53	0.04
Level of assistance from the PES: Light	0.25	0.25	0.25	0.25	0.25	1.00
Level of assistance from the PES: Medium	0.55	0.55	0.55	0.55	0.55	1.00
Level of assistance from the PES: Strong	0.19	0.19	0.19	0.19	0.19	1.00
Married	0.42	0.41	0.43	0.43	0.42	0.04
Max. commuting time (minutes)	23.45	23.82	23.62	23.56	23.35	0.28
No child	0.57	0.58	0.57	0.57	0.57	0.95
Occupation targeted: Agriculture	0.03	0.03	0.03	0.03	0.03	0.99
Occupation targeted: Art and crafts	0.01	0.01	0.01	0.01	0.00	0.98
Occupation targeted: Banking, insurance, real est.	0.01	0.01	0.01	0.01	0.01	0.95
Occupation targeted: Business support services	0.16	0.16	0.16	0.16	0.16	0.90
Occupation targeted: Comm, media, digital	0.02	0.02	0.02	0.02	0.02	0.99
Occupation targeted: Construction, public works	0.07	0.06	0.07	0.06	0.07	1.00
Occupation targeted: Health	0.04	0.04	0.04	0.04	0.04	1.00
Occupation targeted: Industry	0.08	0.08	0.08	0.08	0.08	1.00
Occupation targeted: Maintenance	0.04	0.04	0.04	0.04	0.04	1.00
Occupation targeted: Missing	0.00	0.00	0.00	0.00	0.00	0.56
Occupation targeted: Performing arts	0.02	0.02	0.02	0.02	0.02	0.99
Occupation targeted: Personal services	0.19	0.19	0.19	0.19	0.19	1.00
Occupation targeted: Sales	0.15	0.15	0.15	0.15	0.15	0.98
Occupation targeted: Tourism, leisure	0.09	0.09	0.09	0.09	0.09	0.99
Occupation targeted: Transport	0.10	0.10	0.10	0.10	0.10	0.98
Reservation wage (in euros)	2702.90	2864.36	2799.65	2808.66	2838.10	0.52
Skill level: Higher occupation	0.15	0.15	0.16	0.15	0.15	0.66
Skill level: Intermediate occupation	0.70	0.69	0.69	0.69	0.69	0.85
Skill level: Lower occupation	0.12	0.13	0.13	0.12	0.13	0.93
Skill level: Missing	0.03	0.03	0.03	0.03	0.03	0.18
UB status: Not eligible to UB	0.49	0.50	0.49	0.48	0.49	0.41
UB status: Receives UB	0.51	0.50	0.51	0.52	0.51	0.41
Unemployment duration (in months)	15.29	15.35	15.20	15.43	15.32	0.80
Work experience (in months)	9.71	9.00	9.32	9.57	9.38	0.35
N. Obs.	10099	10092	10108	10094	10102	

Note: Columns (1) to (5) characterize job-seekers by their treatment assignment and report mean values (as a share of the sample unless stated otherwise); column p reports the p-value from the F-test for joint significance of treatment coefficients in the regressions of each covariate on treatment assignment.

A.5 Attrition differential

Table A.2 displays the results of the regression:

$$Y_i = \alpha + \sum_k \beta_k \{T_i = k\} + \epsilon_i$$

among job seekers who received an email, where T_i is job seeker i 's received treatment, and Y_i corresponds to a binary indicator of having completed the survey (rated the top two ads and accessed the final page). The \mathcal{P} -REC treatment serves as the reference category. A F-test of the joint nullity of coefficients associated to \mathcal{U} -REC, MIX^{-1/4}, MIX^{-1/2} and MIX^{-3/4} yields a F-stat 1.885 (p=0.11). Accordingly, we do not attempt to model attrition differential.

Table A.2: Survey completion

Dependent variable	Completed the survey
MIX ^{-1/4}	-0.003 (0.005)
MIX ^{-1/2}	0.009 (0.006)*
MIX ^{-3/4}	0.001 (0.005)
\mathcal{U} -REC	-0.004 (0.005)
Strata fixed effects	Yes
N.Obs.	50 495
Control mean (\mathcal{P} -REC)	0.176

Note: The \mathcal{P} -REC treatment group is used as the reference category. Robust standard errors are in parentheses. *, **, ***: significance at 10%, 5%, and 1%.

B Model appendix

B.1 Proof of propositions

Proof. Proof of Proposition 3.1. The utilities associated with accepting a job offer and with continued search are given by:

$$(1 + rdt)V_e(x, y, \varepsilon_y) = [(U(x, y) + \varepsilon_y)dt + (1 - qdt)V_e(x, y) + qdtV_{0,u}(x)] \quad (\text{B.1})$$

$$\begin{aligned} & (1 + rdt)V_{0,u}(x) \\ & = u(b)dt \\ & + \alpha_0 dt \int_{y, \varepsilon_y: C(x, y, \varepsilon_y)=1} (p(x, y)V_e(x, y, \varepsilon_y) + (1 - p(x, y))(V_{0,u}(x) - R) - k) dF_0(y) dF(\varepsilon_y) \\ & + \alpha_0 dt \int_{y, \varepsilon_y: C(x, y, \varepsilon_y)=0} V_{0,u}(x) dF_0(y) dF(\varepsilon_y) + (1 - \alpha_0 dt)V_{0,u}(x), \end{aligned} \quad (\text{B.2})$$

where $C(x, y, \varepsilon_y) \in \{0, 1\}$ describes the application behavior.

Equation B.1 gives $V_e(x, y, \varepsilon_y) - V_{0,u}(x) = (U(x, y) + \varepsilon_y - rV_{0,u}(x))/(r + q)$, which, plugged in equation B.2 and denoting by $\bar{R} := (r + q)R$, $\bar{k} := (r + q)k$, leads to:

$$\begin{aligned} & rV_{0,u}(x) \\ & = u(b) + \frac{\alpha_0}{r + q} \int_{y, \varepsilon_y: C(x, y, \varepsilon_y)=1} p(x, y) (U(x, y) - U_0^*(x, p(x, y)) + \varepsilon_y) dF_0(y) dF(\varepsilon_y), \end{aligned}$$

the decision to apply is thus

$$C(x, y, \varepsilon_y) = \mathbf{1} \{U(x, y) - U_0^*(x, p(x, y)) > -\varepsilon_y\}.$$

□

Proof of Proposition 3.3. Equation (1.16) is simply obtained plugging the definition of the function Γ^m given in equation (1.12) into the discounted value equation of equation (1.11).

- The optimality of the selection index Γ^m is straightforward. Assume S is another index and consider the associated selection rule $1(S(p, u) > \gamma_s)$ with γ_s the quantile of order $1 - s$ of S . $E(\Gamma^m 1\{S(P, U) > \gamma_s\}) = E(\Gamma^m E(1\{S(P, U) > \gamma_s\} | \Gamma^m))$. Let $g(\Gamma^m) = E(1\{S(P, U) > \gamma_s\} | \Gamma^m)$. We can rewrite $E(\Gamma^m 1\{S(P, U) > \gamma_s\}) = E(\Gamma^m g)$. Lets consider now the supposed optimal selection rule $1(\Gamma^m > \zeta_s)$ with again ζ_s the quantile of order $1 - s$ of Γ^m . We have $E(\Gamma^m g) = E(\Gamma^m g | \Gamma^m > \zeta_s) s + E(\Gamma^m g | \Gamma^m < \zeta_s) (1 - s)$. But $E(\Gamma^m g | \Gamma^m < \zeta_s) (1 - s) < \zeta_s E(g | \Gamma^m < \zeta_s) (1 - s)$ and, given $E(g) = s$,

we have $E(g|\Gamma^m < \zeta_s)(1-s) = sE(1-g|\Gamma^m > \zeta_s)$. Thus $E(\Gamma^m g) - E(\Gamma^m|\Gamma^m > \zeta_s)s < E((\Gamma^m - \zeta_s)(g-1)|\Gamma^m > \zeta_s)s < 0$.

- To show the second result in the proposition, consider a RS based on S and call $g(z) = E(\Gamma^m|S = z)$. We have $E(\Gamma^m) = E(g(z)) = E(g(S)|S > \zeta_s)s + E(g(S)|S < \zeta_s)(1-s)$. Thus $E(\Gamma^m) - E(\Gamma^m|S > \zeta_s) = E(g(S)|S > \zeta_s)s + E(g(S)|S < \zeta_s)(1-s) - E(g(S)|S > \zeta_s)$ and thus $E(\Gamma^m) - E(\Gamma^m|S > \zeta_s) = (1-s)(E(g(S)|S < \zeta_s) - E(g(S)|S > \zeta_s))$ which is negative given g is increasing.
- The last result directly follows as $E(\Gamma^m|\Gamma^m = z) = z$.

□

C Additional figures and tables

Table C.3: Information on offers and job seekers respectively used by the preference-based and machine-learning based RSs

Preference-based		Machine learning	
Job seekers	Offers	Job seekers	Offers
Skills	Skills	Skills (SVD, embedding)	Skill (SVD, embedding)
Diploma	Diploma	Diploma	Diploma
Languages	Languages		
Driver's licence	Driver's licence	Driver's licence	Driver's licence
Experience	Experience	Experience	Experience
Occupation (lv. 3)	Occupation (lv. 3)	Occupation (lv. 1, 2, 3)	Occupation (lv. 1, 2, 3)
Working hours	Working hours	Working hours	Working hours
Wage	Wage	Wage (several measures)	Wage (upper, lower bounds)
Location	Location	Location	Location
Geo. mobility		Geo. mobility	
Contract type	Contract type	Contract type	Contract type
		Qualification	Qualification
		Soft skills	Soft skills
			Job description (text)
			Firm description (text)
			Contract type
			Contract duration
			Establishment size
			Establishment status
			Num. applications (ad)
			Num. applications (establishment)
			Num. days since posted
			Geo. soc.-dem. features
		Former occupation	
		Sex	
		Num. children	
		Search obligations	
		Job search type	
		Min. allowance status	
		Days unemployed	
		Age	
		Num. applications	
		Geo. soc.-dem. features	

Table C.4: Descriptive statistics on the populations with clicks

	All	Not Movers	Movers
Q25 Nb clicks	1	1	6
Median Nb clicks	3	3	12
Q75 Nb clicks	8	7	27
Application rate	0.033	0.015	0.206
Average hiring prob.	0.038	0.038	0.040
Average utility on jobs	0.291	0.286	0.336
Unemp. duration (months)	11.4	11.4	10.9
Full time	0.951	0.949	0.972
Long term	0.868	0.867	0.876
Reservation wage (euros/year)	19,845	19,869	19,610
At least Master	0.015	0.015	0.018
More than Bachelor	0.039	0.039	0.042
Bachelor	0.090	0.091	0.081
High school degree	0.229	0.228	0.240
No high school degree	0.494	0.494	0.498
Experience	6.078	6.074	6.118
Nb children	0.930	0.934	0.890
Nb observations	8,105	7,348	757

Notes: The different statistics are presented on all the population where we observe clicks during the four weeks of the test set (“All”), on the population of “Movers” which have at least two clicks and for which we observe either at least one application or one rejection, and on “Not Movers” which are the complementary. Estimation of the binary choice fixed effects model can only be performed on the movers.

Table C.5: Estimates of the model of application on job postings

	(Logit on Full sample)		(FE-Logit on “Movers”)		(FE-Logit on “Movers”)	
	Estimate	Std. error	Estimate	Std. error	Estimate	Std. error
Utility score $\mathcal{U}(i, j)$ (α)	1.374**	0.180			1.101**	0.155
Unif. Utility score $\bar{U}(i, j)$ (α)			0.830***	0.137		
- Inverse of $\mathcal{P}(i, j)$ (β)	0.019**	0.007	0.029***	0.004	0.028**	0.004
Intercept / Avg. indiv. FE	-3.375	0.162	-1.388	0.047	-1.356	0.046
Nb. observations	70,557		17,865		17,865	

Estimation of equation (1.19) modeling applications as logit model.

Notes: Our sample is the set of all applications for job seekers in the transportation and logistic sector during week 44 of 2019, leading to a hiring or not. Fixed effect estimation keeps 70,557 observations for 8,105 job seekers, and 869 of them applying at least once, and 757 “Movers” with one application and 1 click without application. In this latter case, 17,865 observations are kept for estimation. Results are robust to the different negative sampling strategies we considered. Significance levels: < 1% : **, < 5% : *, < 10% : †.

Table C.6: Average hiring rate of the different RSs

	Average over			std (4)
	All (1)	bottom 10% (2)	top 10% (3)	
Hiring rate: θ (x100)				
Benchmark	0.88	0.07	2.76	0.82
\mathcal{U}	0.43	0.03	1.26	0.52
\mathcal{P}	1.52	0.14	4.06	1.34
Δ	1.45	0.15	3.61	1.20
Optimal	1.60	0.16	4.14	1.35
Pseudo	1.59	0.16	4.08	1.33

- (1) Average over the full sample
(2) (resp. (3)) Average over these in the bottom (resp top) 10% of the distribution of the expected value
(4) standard deviation over the full sample

Table C.7: Comparison of the hiring rate of the different RSs

	All (1)	Mean effect over				Std (6)
		bottom 10% Optimal (2)	top 10% Benchmark effect (3)	bottom 10% Benchmark (4)	top 10% (5)	
Hiring rate: θ (x100)						
Optimal - Benchmark	0.73	0.06	2.33	0.09	1.38	0.71
\mathcal{P} - Benchmark	0.64	0.04	2.26	0.07	1.30	0.71
Δ - Benchmark	0.57	0.05	1.89	0.08	0.85	0.61
\mathcal{U} - Benchmark	-0.45	-0.08	-0.71	-0.04	-1.51	0.59
Pseudo - Benchmark	0.72	0.06	2.29	0.09	1.32	0.70
Optimal - \mathcal{P}	0.09	0.02	0.07	0.02	0.08	0.10
Optimal - Δ	0.16	0.01	0.45	0.01	0.53	0.24
Optimal - \mathcal{U}	1.18	0.14	3.04	0.13	2.89	1.04
Optimal - Pseudo	0.01	0.00	0.04	0.00	0.07	0.04

- (1) Average Effect over the full sample
(2) (resp. (3)) Average Effect over these in the bottom (resp top) 10% of the distribution of the effect of the effect on expected value of switching from the benchmark to the optimal RS
(4) (resp. (5)) Average Effect over these in the bottom (resp top) 10% of the distribution of the Benchmark of the expected value
(6) Standard deviation

Table C.8: Shares of clicks and recommendations inside the declared occupational search area

	Clicks	Reco. Γ	Reco. \mathcal{P}	Reco. \mathcal{U}	Applications
Share in occup. search area	0.30	0.55	0.32	0.65	0.37

Columns Reco. Γ , Reco. \mathcal{P} , Reco. \mathcal{U} are computed on the top k best recommendations, where k is the number of observed clicks of individual i .

Chapter 2

Personalization Pitfalls: The Unintended Effects of Using Stated Preference Data in Job Recommendations

This chapter is based on a joint work with Abhijit Banerjee (MIT), Bruno Crépon (CREST-ENSAE), and Cécile Welter-Médée (Insee) ¹.

Abstract: This study examines the effects of incorporating job seekers' stated preferences about job attributes into the job recommender system of the French Public Employment Service (PES). In a randomized controlled trial, job seekers distributed 15 points across five job attributes—occupation, wage, commuting distance, working hours, and contract type—to indicate their priorities. Personalizing recommendations based on these preferences led to a 13.8% increase in clicks and a 23.1% increase in applications compared to the non-personalized algorithm, driven largely by new job ads that job seekers would not have seen without personalization. However, compared to the status quo algorithm without preference collection, the effort involved in the preference elicitation process led to a 26% reduction in clicks on recommendations and a 30% decline in applications to recommendations. These findings suggest that while personalization can improve recommendation quality, the effort required may reduce user interaction, highlighting the need to balance personalization benefits with the drawbacks of increased user effort.

¹The research is the result of a collaboration with France Travail, the Public Employment Service in France. We thank H el ene Caillol, Peggy Duhayon, Thierry Foltier, Pascale Marquoin, Paul Mariage, Cyril Nouveau, Michael Paulus, Camille Qu er e, Sideline Taing and Chantal Vessereau for their operational support. We are grateful to Juliette Lynch for her outstanding research assistance. The authors retained full intellectual freedom throughout this process, and any errors are our own. This project received IRB approvals from MIT (2112000531) and the Paris School of Economics (2021-027).

Introduction

Addressing the multifaceted nature of unemployment requires a close examination of the persistent matching frictions within the labor market. These frictions create a paradoxical situation where job vacancies coexist with an available workforce, leading to prolonged unemployment spells and inefficient utilization of human resources. Such frictions arise from informational asymmetries, geographic mismatches, and the complexities of aligning workers' skills with job requirements.

As a response to these challenges, Public Employment Services (PES) have turned to algorithmic solutions to better match job seekers with vacancies. Research has shown that automated job recommendations can lead to improvements in interview rates [Bélot et al., 2018, Li et al., 2020] and long-term employment outcomes [Bélot et al., 2022, Altmann et al., 2022, Behaghel et al., 2024], although some studies report only modest effects [Le Barbanchon et al., 2023]. Among the different types of job recommendation algorithms used, expert systems or “knowledge-based” systems are a notable approach, accounting for about 12.7% of job recommender systems in recent studies [Freire and de Castro, 2021] and are widely implemented in PES globally [World Bank, 2023]. Unlike machine learning models that rely on data-driven inferences, expert systems operate based on explicit, predefined rules crafted by human experts. These “if-then” rules consider a multitude of factors related to both job seekers and job vacancies, offering a structured way to address mismatches in the labor market.

An illustrative example of such an expert system is the ELISE Search and Match platform developed by WCC, a Dutch technology company. The ELISE platform is utilized by various European countries—including France, Luxembourg, Austria, and Germany—in their public employment services to match job seekers with employers [World Bank, 2023]. In France, this matching system plays a central role in the operations of the PES. It is integrated into the PES online platform used by job seekers, serving both as a search engine and a recommender system. Recruiters also use the same platform to find suitable candidates for their vacancies. Additionally, PES caseworkers utilize this matching algorithm to identify suitable candidates for positions where employers have specifically requested assistance in finding appropriate candidates.

The matching process begins by collecting detailed information from job seekers, such as their qualifications, work experience, skills, and specific search criteria—including desired occupation, reservation wage, preferred location, contract type, and working hours. Job vacancies are similarly detailed, encompassing factors like job requirements, geographical location, salary, contract type, and working hours. The expert system applies its predefined rules to assess the compatibility between a job seeker's profile and each job

vacancy. For instance, one rule might verify if the offered salary meets the job seeker's reservation wage, while another evaluates whether the job's location is within a reasonable commuting distance for the job seeker. To balance the various factors involved in matching, each criterion is assigned a weight that reflects its importance in the overall process. These weights are determined by human experts who leverage their understanding of labor market dynamics to prioritize certain attributes over others. By calculating a matching score for every job seeker–vacancy pair—where higher scores indicate better matches—the system identifies the most suitable job opportunities for each individual. The vacancies with the highest matching scores are then recommended to job seekers, aiming to efficiently connect them with positions that best meet their needs and qualifications.

Despite these advancements, a significant limitation remains in the uniform application of attribute weights, leading to a lack of personalization. The French PES matching system applies the same weights to each attribute across all job seekers, failing to account for the varied priorities of individuals. Indeed, several studies highlight heterogeneity in workers' preferences for job attributes. Workers value various job characteristics differently, and these preferences can significantly influence job choices and contribute to labor market outcomes. For instance, [Wiswall and Zafar \[2017\]](#) explore how gender differences in workplace preferences affect job choices, human capital investments, and contribute to the gender wage gap. Similarly, [Mas and Pallais \[2017b\]](#) examine workers' valuation of flexible scheduling and work-from-home options, finding that women, especially those with children, have a higher willingness to pay for these attributes. Using French administrative data [Le Barbanchon et al. \[2021\]](#) finds that unemployed women value shorter commutes around 20% more than men, resulting in a willingness to accept lower wages. [Maestas et al. \[2023\]](#) investigate workers' willingness to trade wages for better job characteristics, finding that preferences vary across demographic groups and contribute to wage inequality. These findings underscore the importance of incorporating individual preferences into job matching algorithms to better cater to diverse job seeker priorities. Additionally, [Banerjee and Chiplunkar \[2024\]](#) found that Indian job seekers prioritize salary, location, and job title the most. Notably, the study reveals that job location holds greater importance for women than men, reflecting a preference for jobs closer to their homes.

Incorporating user input into these systems could address the issue of personalization. For instance, [Banerjee and Chiplunkar \[2024\]](#) showed that a more tailored matching system, which considers diverse priorities and preferences of job seekers, could potentially enhance job seeker satisfaction and employment outcomes. For quite some time, researchers in computer science have recognized that the effectiveness of recommender systems extends beyond mere accuracy [[Swearingen and Sinha, 2001](#)], underscoring the sig-

nificance of user involvement in the recommendation process. Incorporating individual preferences and feedback leads to more precise and relevant suggestions, enhancing not only the system's accuracy but also its relevancy [Pu et al., 2012, Parra and Brusilovsky, 2015]. For instance, systems like "TasteWeights" enable users to fine-tune weights on different parameters, which not only improves the accuracy of music recommendations but also significantly enhances user engagement and satisfaction [Bostandjiev et al., 2012]. In the context of job search, Bächli et al. [2024] propose a method that allows job seekers to fine-tune the parameters of an occupational recommender system. These parameters (α and β) enable job seekers to either target their own skill profile or previous occupation (α), and anchor suggestions to other requirements in their previous occupation (β).

However, involving users in the recommendation process introduces a trade-off between the recommendations' accuracy and user effort. While eliciting detailed preferences can enhance recommendation precision, it may also increase cognitive load and user burden, potentially reducing user engagement [Pu et al., 2012]. One explanation from behavioral decision theories is that individuals tend to prioritize minimizing cognitive effort over maximizing accuracy because the effort is immediately perceptible, whereas the benefits of increased accuracy are delayed and uncertain [Häubl and Trifts, 2000]. Furthermore, studies have shown that while collecting user input initially enhances recommendation precision, there are diminishing returns to additional user effort, with the most substantial gains in recommendation quality often achieved with only minimal input (e.g. Drenner et al. [2008]). This underscores the importance of balancing accuracy with user effort to maintain user engagement.

Therefore, selecting an appropriate method for eliciting user preferences is essential to balance personalization accuracy and user engagement. Preference elicitation methods are generally classified into revealed and stated preference approaches. Revealed preferences infer choices from observed behaviors but are limited to existing options and cannot assess hypothetical scenarios. Stated preferences directly solicit individual preferences, allowing for hypothetical contexts but may face concerns about hypothetical bias and external validity [Diamond and Hausman, 1994, Manski et al., 2000]. Interestingly, Banerjee and Chiplunkar [2024] found that job seekers' reported preferences regarding job aspirations and priorities were consistent regardless of whether incentives were provided to elicit their *true* preferences. This consistency suggests that the preferences elicited were genuine rather than strategic responses aimed at increasing their chances of securing a job.

Within stated preference methods, there are two main approaches: compositional and decompositional [Helm et al., 2004, Weernink et al., 2014, Marsh et al., 2016]. Compositional methods directly ask individuals to allocate importance to different job attributes,

while decompositional methods, such as discrete choice experiments (DCE), infer preferences based on the choices individuals make between different job options. While DCE has gained popularity among labor economists as an alternative to revealed preference methods for estimating compensating wage differentials [Mas and Pallais, 2017a, Wiswall and Zafar, 2017], our study employs a compositional approach to capture job seekers' preferences more directly. Common compositional techniques include direct rating, direct ranking, and point allocation (for a review, see Van Ittersum et al. [2007], Zardari et al. [2014]). However, direct rating and ranking methods often lack differentiation among attributes due to the absence of trade-offs, leading participants to rate all attributes as highly important [Krosnick and Alwin, 1988, McCarty and Shrum, 2000]. The point allocation method addresses this issue by requiring individuals to distribute a fixed number of points among attributes, introducing explicit trade-offs and encouraging careful consideration of attribute importance. Although the direct-rating method shows higher test-retest reliability due to its simplicity [Bottomley et al., 2000], this does not necessarily indicate greater validity in capturing true preferences [Bottomley and Doyle, 2013].

Given our objective to derive precise attribute weights for personalizing job recommendations, we use the point-allocation method in our study. This approach effectively engages respondents in trade-offs that reflect their true preferences without imposing excessive cognitive burdens. We allocate 15 points to participants, asking them to distribute these among five job attributes: occupation, commuting distance, wage, contract type, and working hours. This method allows job seekers to indicate the relative importance of each attribute, providing data that can be directly applied as weights in our personalized job matching algorithm.

With this perspective, our research aims to explore the impact of integrating these individual job seekers' preferences into the PES matching algorithm on recommendation appreciation and adoption. We also aim to understand the variations in these preferences according to socio-demographic factors and unemployment history. To this end, we have implemented an intervention involving a personalized version of the PES matching algorithm that recommends job vacancies to job seekers based on their preferences regarding the importance of the following job attributes: occupation, wage, distance from home, working hours, and type of contract. This intervention is facilitated through a web interface designed to collect job seekers' preferences about job attribute importance and subsequently offer them job vacancies that align with these personalized preferences. Our study is divided into two sequential phases. In the first phase, we conduct a pilot study to test and verify the reliability of the point-allocation method for collecting job attribute preferences. Participants in this phase are asked to distribute a fixed number of points among key job attributes (occupation, salary, commuting distance, type of contract,

working hours) to capture the relative importance of each attribute. In the second phase, a large-scale randomized controlled trial (RCT) is implemented, where participants state their preferences using the same method as in the pilot study. The RCT involves three arms: one third of job seekers (henceforth the “full treatment” group) receiving recommendations tailored to their personalized weights, one third (“partial treatment” group) providing preferences but receiving standard algorithm recommendations, and one third (control group) following the existing PES algorithm without preference input. This design allows us to assess the impact of preference-based personalization on job seeker engagement. The pilot experiment was conducted on approximately 20,000 job seekers in March 2022, before launching a full-scale randomized controlled trial on 250,000 job seekers in June, 2022.

The pilot study revealed that job seekers generally engaged thoughtfully with the preference elicitation tasks, indicating the reliability of the point-allocation method used to capture job attribute preferences. In the large-scale experiment, job seekers’ responses to the point allocation task showed a considerable divergence from the standard weights used in the PES matching algorithm, with preferences displaying a more balanced distribution across attributes. Notably, wage and commuting distance were rated higher than occupation, which contrasts with the emphasis placed on occupation by the standard PES system.

Our findings reveal that integrating personalized preferences into the recommender system resulted in an increased number of job recommendations. This occurred because, with the standard weights used in the PES algorithm, many job ads did not accumulate sufficient scores to be recommended to job seekers. By incorporating job seekers’ own point allocations, more ads met the criteria for recommendation. Consequently, while the number of recommendations increased, some of these ads included positions that were less aligned with job seekers’ preferred occupations and salary expectations.

Then, we examine interactions between job seekers and the job recommendations made during the experiment (clicks, intentions to apply and actual applications). We do not look at organic clicks and applications on the PES platform. To isolate the effect of personalization, we compared job seekers in the full treatment group (receiving personalized recommendations) with those in the partial treatment group (receiving standard recommendations). We find that personalization alone showed significant and positive effects : job seekers in the full treatment group clicked on recommendations 13.8% more, exhibited a 20.8% increase in intentions to apply, and a 23.1% rise in actual applications compared to the partial treatment group. These effects were primarily driven by the new job ads introduced through the personalized algorithm rather than a mere reallocation of interest from standard recommendations.

However, these positive outcomes were not enough to counterbalance the effort expenditure made by treated individuals when completing the preference elicitation task. When comparing the full treatment group to the control group, we found a 26% reduction in clicks, a 27% decline in intentions to apply, and a 30% drop in actual applications. Additionally, platform visits decreased by 24%, indicating reduced overall engagement. Further, we examined the impact of the point allocation task itself by comparing the partial treatment group to the control group. Here, we observed a sharp decline in engagement: clicks decreased by 35%, intentions to apply fell by 40%, and applications dropped by 43%. Platform visits also declined by 25.7%. This suggests that the point allocation task set higher expectations for job recommendation accuracy. When these heightened expectations were not met, job seekers experienced disappointment, leading to reduced engagement. Overall, our results align with previous research on the trade-off between accuracy and effort in recommender systems [Pu et al., 2012]. They highlight that the benefits of personalization must be weighed against the potential for increased user burden, emphasizing the need for careful design in the implementation of preference-based algorithms.

The findings have important implications for the design of job recommender systems. Personalized recommendations lead to better engagement; hence, policymakers and platform designers might consider incorporating more user-driven features into these systems. However, the benefits of personalization must be weighed against the potential for increased user burden. Balancing personalization with user effort seems essential to maximize the effectiveness of recommender systems in the labor market.

This paper contributes to several streams of literature. First, we contribute to the research on the impact of recommender systems on the labor market by providing empirical evidence on how personalized job recommendations affect job seeker engagement and recommendation adoption within a PES context. Our findings add to the evidence on the effectiveness of job recommender systems [Bélot et al., 2018, Li et al., 2020, Bélot et al., 2022, Altmann et al., 2022, Behaghel et al., 2024, Le Barbanchon et al., 2023]. Second, we contribute to the literature on job attribute preferences by analyzing how job seekers' preferences vary according to socio-demographic factors and unemployment history, aligning with previous studies that highlight heterogeneity in workers' valuation of job attributes [Wiswall and Zafar, 2017, Mas and Pallais, 2017b, Le Barbanchon et al., 2021, Feld et al., 2022, Maestas et al., 2023]. Third, we contribute to the literature on different types of recommender systems in job matching by evaluating a personalized expert system that incorporates user input [Freire and de Castro, 2021].

The remainder of this paper is structured as follows. Section 2 describes the institutional context and the study design, including the technical challenges encountered. Sec-

tion 3 describes the data we use. Section 4 presents the results of the pilot study, which evaluated the reliability of the preference elicitation method. Section 5 analyzes the stated preferences for job attributes collected during the experiment. Section 6 provides a descriptive analysis of the recommendations generated by both the standard and personalized algorithms. Section 7 evaluates the results of the large-scale experiment, namely the impact of weight personalization and the preference elicitation task on recommendations appreciation and adoption.

1 Context and study design

1.1 Institutional context and the PES's current job recommender system

The French Public Employment Service (PES) utilizes an expert system to match job seekers with suitable vacancies based on their skills and preferences. This system is a specialized application of artificial intelligence designed to replicate the decision-making processes of a human expert in job matching. Unlike machine learning models, which rely on patterns identified from historical data to make predictions, expert systems like the one employed by the PES use explicit, pre-determined rules created by human experts to guide their recommendations.

The PES's job matching process is based on the ELISE platform, a system developed by the Dutch company WCC. This platform is used by several European countries, such as Luxembourg, Austria, and Germany, as well as in non-European regions like Singapore and Saudi Arabia [World Bank, 2023]. The ELISE algorithm works by converting the characteristics of job seekers and job postings into structured data points, facilitating the matching process. The algorithm evaluates ten key criteria, divided into two broad categories: (i) job seeker profiles and (ii) job attributes. The first category includes aspects such as work experience, educational background, and skills, while the second covers job-specific factors like occupation, location, and salary. For example, the algorithm assesses how well the advertised salary of a job aligns with a job seeker's reservation wage and how closely the required experience matches the job seeker's actual work history. This structured approach ensures a thorough assessment of both the job seeker's suitability for a job and the job's relevance to the job seeker.

To provide a clearer understanding of the matching process, the following table lists the criteria used in the PES algorithm along with their respective weights, highlighting the relative importance of each factor in generating job matches.

(1) Profile	Weight w_k^{prof}	(2) Job attributes	Weight w_k^{attr}
Skills (hard and soft)	10.7	Occupation	10.7
Education	1.1	Working hours	1.1
Languages	1.1	Reservation wage	2.1
Driving license	1.1	Commuting distance	1.1
Working experience	1.1	Duration and type of contract	0.1

Table 2.1: Criteria and weights used in the standard PES matching algorithm

In the following, we describe how this expert matching system is used to recommend vacancies to job seekers at the PES. For each criterion k enumerated in Table 2.1, is defined a quantitative adequacy measure $a_k(i, j) \in [0, 1]$. This measure represents the degree of compatibility between the job seeker i 's preferences and the recruiter j 's requirements pertaining to criterion k . As an illustration, the adequacy measure for wage would equate to 1 if the salary proposed in the job vacancy exceeds the individual's reservation wage. Conversely, in a scenario where the proposed salary is lower than the reservation wage, the adequacy measure would yield a value between 0 and 1, decreasing proportionally as the proposed wage diverges from the reservation wage. The computation methodology for each adequacy measure is described in the appendix (refer to section A).

Subsequently, the matching score between a job vacancy and a job seeker is calculated as the weighted sum of each individual adequacy measure, considering all 10 criteria for both the vacancy and the job seeker. Importantly, the weights applied in this algorithm are uniform across the entire population and are enumerated in Table 2.1. Formally, the *standard* matching score between a job seeker i and a vacancy j , denoted as $s^s(i, j) \in [0, 1]$, is computed as follows:

$$s^s(i, j) = f_{ij} \times \left[\sum_{k=1}^5 \frac{w_k^{prof}}{\sum_{k=1}^5 w_k^{prof}} a_k^{prof}(i, j) + \sum_{k=1}^5 \frac{w_k^{attr}}{\sum_{k=1}^5 w_k^{attr}} a_k^{attr}(i, j) \right] \quad (2.1)$$

with:

- w_k^{prof} the weight assigned to the *profile feature* k ,
- $a_k^{prof}(i, j)$ the adequacy measure between the recruiter's requirement regarding the profile feature k and the job seeker's profile characteristic k ,
- w_k^{attr} the weight assigned to the *job attribute* k ,
- $a_k^{attr}(i, j)$ the adequacy measure between the vacancy's attribute k and the job seeker's preference regarding attribute k ,

- $f_{ij} \in \{0, 1\}$ a filter variable which is equal to 0 if at least one of the following adequacy measure is null: commuting distance, languages and driving license.

Given the set of available vacancies, denoted as \mathcal{F} , the system generates a ranking of vacancies for each job seeker based on the computed matching score from Equation 2.1. This ranking can then be used to recommend the most appropriate vacancies tailored to the individual job seeker's needs and qualifications. The vacancies constituting \mathcal{F} are those posted directly on the PES online platform and are available when the job seeker is actively seeking employment¹.

For any given job seeker i , we designate the set of candidate vacancies as \mathcal{J}_i^s . This set comprises job advertisements from the available vacancies \mathcal{F} , each of which has garnered a matching score equal to or greater than 50%. Formally, this set is defined as follows:

$$\mathcal{J}_i^s = \{j \in \mathcal{F} \mid s^s(i, j) \geq 0.5\} \quad (2.2)$$

There may be instances where \mathcal{J}_i^s is an empty set, indicating that no job vacancy has achieved a satisfactory matching score.

We further introduce R_{ij}^s , which indicates the ranked position of a vacancy j within the sequence of candidate vacancies when ordered according to their standard matching scores with job seeker i . It is calculated as follows:

$$R_{ij}^s = \sum_{k \in \mathcal{J}_i^s} \mathbb{1}(s^s(i, j) < s^s(i, k)) + 1 \quad (2.3)$$

with $\mathbb{1}(x)$ an indicator function. Finally, the set of recommendations $\mathcal{R}_s(i)$ made to job seeker i is defined as:

$$\mathcal{R}_s(i) = \{j \in \mathcal{J}_i^s \mid R_{ij}^s \leq 150\} \quad (2.4)$$

As the definition implies, job advertisements with the highest scores within the set \mathcal{J}_i^s are recommended to job seeker i , with a cap at a maximum of 150 ads. In practice, the system presents vacancies to the job seeker in descending order of matching scores. In the event where the set of candidate vacancies \mathcal{J}_i^s is empty, no job advertisement is recommended to the job seeker.

¹It is important to note that the market share of PES job vacancies among all online job vacancies varies significantly across different occupation types. For instance, recruiters for less-skilled and blue-collar occupations tend to prefer the PES platform for their recruitment needs. Conversely, recruiters looking to fill skilled white-collar positions typically gravitate towards other job boards.

1.2 Study design

1.2.1 Objectives of the study and intervention

While the PES expert system increases the efficiency of the matching process, it also has several notable drawbacks. A primary limitation is the lack of personalization: the system applies uniform weights and matching rules to all individuals, failing to account for the unique preferences and circumstances of individual job seekers and vacancies. This generic approach does not adequately account for the different priorities job seekers may have, as fixed weights heavily emphasize targeted occupations while minimally considering factors such as salary, commuting distance, working hours, and contract type. In addition, the reliance on rigid, predefined rules may not accurately reflect the complexities of the labor market, potentially making the system opaque and reducing trust among job seekers. Finally, the effectiveness of the system depends heavily on the quality of the input data, particularly the accuracy and completeness of job seeker profiles and vacancy details.

Our research aims to address these limitations by examining how incorporating individual job seekers' preferences for prioritizing job search criteria affects their engagement with recommendations. Specifically, we focus on five search criteria: salary, occupation, commuting distance, working hours, and contract type, as detailed in Table 2.1. Importantly, our study does not aim to modify the absolute values of job seekers' search criteria, but instead treats them as given². In addition, we aim to examine how these preferences for the importance of job attributes vary across different sociodemographic groups and unemployment histories. Achieving these research goals requires collecting and integrating job seekers' priorities for the five job attributes mentioned above, allowing the system to better reflect their unique priorities.

Importantly, we are not interested in collecting and integrating into the system the importance of attributes related to job seeker profiles, such as skills or education. These attributes are inherently linked to the requirements of recruiters and are important for meeting the needs of employers. They are more about job qualifications than personal preferences, making them unsuitable for the kind of personalization we want to implement.

In the following, we explain the methodology for incorporating job seekers' individual preferences into the standard PES matching algorithm. A straightforward approach to tailoring this standard algorithm to accommodate personal preferences regarding job

²On registering with the PES, job seekers formalize these criteria in a contract with their caseworker. Jobseekers are required to actively pursue opportunities that match these criteria, and sanctions, including deregistration and benefit suspension, are applied after unjustified refusals of job offers that match these criteria.

search attributes is to replace the fixed weights associated with job attributes (as denoted in equation 2.1) with individual-specific weights. To this end, we introduce the concept of the personalized matching score, defined as:

$$s^p(i, j) = f_{ij} \times \left[\sum_{k=1}^5 \frac{w_k^{prof}}{\sum_{k=1}^5 w_k^{prof}} a_k^{prof}(i, j) + \sum_{k=1}^5 \frac{w_{i,k}^{attr}}{\sum_{k=1}^5 w_{i,k}^{attr}} a_k^{attr}(i, j) \right] \quad (2.5)$$

where $w_{i,k}^{attr}$ denotes the weight that *job seeker i* assigns to the *job attribute k*.

Utilizing this personalized matching score, we can generate recommendations for job seekers in a manner similar to the previous method. We define \mathcal{J}_i^p as the set of candidate job advertisements that the personalized algorithm can recommend to the job seeker *i*:

$$\mathcal{J}_i^p = \{j \in \mathcal{F} \mid s^p(i, j) \geq 0.5\} \quad (2.6)$$

The personalized matching algorithm recommends to job seeker *i* the following set of vacancies, which are ranked according to their personalized matching score:

$$\mathcal{R}_p(i) = \{j \in \mathcal{J}_i^p \mid R_{ij}^p \leq 150\} \quad (2.7)$$

Here, R_{ij}^p represents the position of the vacancy *j* in the ranking of candidate vacancies according to their personalized matching score with *i*. This personalized approach allows for a more nuanced and individually-tailored job matching process that aligns more closely with the job seeker's unique preferences.

The intervention we implemented and that we examine in this paper involves recommending job vacancies to seekers based on the personalized matching algorithm. This intervention is facilitated through a web interface, designed to collect the preferences of job seekers, convert these preferences into weights utilized in Equation 2.5, and subsequently offer them job vacancies that align with these personalized preferences.

Our intervention entails the incorporation of job seekers' preferences regarding the importance of various job attributes as a critical input in the PES matching system. The precise determination of the importance of these job attributes is a critical yet complex endeavor. We need to ensure that the collected preferences are accurately depicted and suitable for application as weights in the matching algorithm, all while ensuring the preference elicitation process does not impose an excessive burden on job seekers. Given this complexity, our research unfolds in two distinct but sequential phases:

1. A pilot study: the primary objective of this pilot study is to scrutinize the relevance and reliability of our proposed method for eliciting preferences. This includes evaluating the effectiveness of our preference collection methods, verifying the accuracy of

the captured data, and determining the feasibility of the method when scaled up.

2. A large-scale experiment: the aim of this Randomized Controlled Trial (RCT) is to examine the potential impact of personalizing the recommendation algorithm to align with individual job seekers' preferences. Specifically, we aim to assess whether a personalized approach to job matching can improve recommendations appreciation and adoption by job seekers, whether it be through higher click rates or application rates. The results of this trial will provide valuable insights into the effectiveness of personalized job matching systems and guide future developments in this area.

Prior to detailing the online platform utilized in the two phases of our study (refer to Subsection 1.3), we outline the experimental design of both the pilot study and the large-scale experiment in the forthcoming subsections.

1.2.2 Measuring job attribute importance

The intervention we are implementing in our paper involves measuring the importance that each job seeker places on various job attributes, such as salary, commuting time, occupation, working hours, and type of contract. However, this type of preference data is not available at the PES: job seekers are only asked about the values of their job search criteria, such as their reservation wage or maximum commuting distance, but not about their relative importance. Fortunately, there are numerous methods for collecting data on the importance of product or service attributes, many of which have a long history in various fields, particularly in marketing research. In this section, we describe the preference elicitation method we have selected and explain why we chose it.

We opted for a stated preference elicitation method to measure attribute importance, rather than relying on revealed preference data. This involves collecting job seekers' preferences in a controlled experimental setting, rather than through their actual labor market behavior. Several reasons explain this choice. First, job attributes often have limited variability in the real labor market. Indeed, the vacancies to which a job seeker is exposed on the PES website and among which they can choose (to click or to apply) are determined by the individual's (often narrow) queries on the PES search engine. Second, job attributes of vacancies can be too closely correlated in the real labor market, e.g., certain occupations are inherently part-time or offered on a temporary contract basis, which makes revealed preference data unreliable for determining the importance of each job attribute taken separately. Third, web data on clicks and applications take time to collect and are not exhaustive, as job seekers may also use other job search channels besides the PES job board.

Stated preference methods can be used to collect importance weights and can be categorized as either compositional or decompositional approaches [Helm et al., 2004, Weernink et al., 2014, Marsh et al., 2016]. Compositional approaches require individuals to directly state the importance of each attribute of an item using a predefined importance scale. On the other hand, decompositional approaches, such as Conjoint Analysis (CA) (i.e., Discrete Choice Experiments), try to infer individuals' preferences by presenting them with hypothetical multiattribute items (products, services, jobs, etc.) that vary in terms of their attributes and asking them to rate or rank alternatives, or choose their preferred one. Although decompositional methods have the advantage of inferring an attribute's importance even if individuals are not consciously aware of it, they are unsuitable for our case for several reasons. First, CA methods require defining a finite number of levels—typically fewer than five—for each attribute to construct the different alternatives presented to individuals. Setting appropriate levels for each job attribute is challenging. For the salary attribute, we can straightforwardly vary levels around the reservation wage desired by the individual. However, setting levels for the occupation attribute is more problematic. Occupations differ vastly, and there is no universally accepted or quantifiable measure to determine the distance or difference between them, making it difficult to create meaningful variations. Second, estimating individual-level importance weights from CA data requires collecting a large number of data points per individual through extensive choice tasks, which is impractical due to potential respondent fatigue and resource constraints. Therefore, decompositional methods may not be feasible for our purposes, and alternative approaches that directly elicit attribute importance may be more appropriate.

Various techniques exist within compositional stated preference methods to collect attribute weights [Ryan et al., 2001, Van Ittersum et al., 2007, Zardari et al., 2014]. Common methods include direct-rating, direct-ranking, and point-allocation (or constant-sum) techniques. The *direct-rating method* asks individuals to rate each attribute on a scale (e.g., 1 = “unimportant” to 7 = “important”); however, without trade-offs, participants may rate all attributes as highly important, leading to a lack of differentiation [Krosnick and Alwin, 1988, McCarty and Shrum, 2000]. Additionally, individuals may interpret the rating scale differently due to cultural differences or ambiguities in scale labels, complicating comparisons between respondents. The *direct-ranking method* requires participants to rank-order attributes, capturing preference order but not the magnitude of differences. While it prompts individuals to consider trade-offs and can provide sufficient variance, it does not directly yield weights; deriving weights from rankings necessitates assumptions about equal importance differences between ranks. The *point-allocation method*, also known as the constant-sum method, involves distributing a fixed number of points (e.g.,

100) among attributes, introducing explicit trade-offs since assigning more points to one attribute reduces the points available for others [Doyle et al., 1997, Bottomley and Doyle, 2013]. This approach encourages respondents to be more attentive to the importance they assign to each attribute and prevents the common issue of uniformly high importance ratings across all attributes. Although it may involve more cognitive effort than direct rating—requiring individuals to keep track of allocated and remaining points—it is not overly demanding and is simple to implement. While the direct-rating method shows higher test-retest reliability due to its simplicity [Bottomley et al., 2000], this does not necessarily indicate greater validity in capturing true preferences [Bottomley and Doyle, 2013]. Despite the differences among these methods, the literature remains inconclusive regarding the most appropriate approach to accurately capture individual judgments.

In our research, we chose to implement the point-allocation technique to collect preferences for job attributes. This method requires individuals to make trade-offs between different attributes, ensuring sufficient discrimination and facilitating easy comparison across individuals. It provides importance weights without any particular assumptions, which can be directly used in the matching algorithm. While the rating method may not generate enough variance across attributes and the ranking method does not directly provide weights, the constant-sum method balances cognitive demand and simplicity, making it suitable for our study.

1.2.3 Pilot study: experimental design and sampling frame

Experimental design The goal of the pilot study is to test the effectiveness and the reliability of the constant-sum question as a way to elicit job attribute importance, before using it our large-scale experiment. Specifically, we implement and compare in the pilot study two preference elicitation methods: the constant-sum method (as described in the previous section) as well as a decompositional inspired question consisting in classifying vacancies which have been posted on the PES website. First, we implement the constant-sum method in order to collect individuals' job attribute importance (regarding salary, commuting time, occupation, working hours and type of contract) and then we compare the data obtained with the results obtained from the ranking question. This protocol serves a dual purpose: firstly, to check if preferences revealed through ranking are consistent with those obtained through the constant-sum question; and secondly, to assess how accurately the weights obtained from the constant-sum question reflect revealed preferences. In practice, after administering the constant-sum question and having collected individuals' job attribute importance, we ask pilot survey participants to rank a few real job ads (and not artificially created ones as in classical decompositional approaches) that were posted on the PES website. The set of vacancies to rank are generated by the person-

alized version of the PES matching algorithm, i.e. using the individual's weights collected through the constant-sum question. Vacancies to rank are shown to the respondent in a random order and are described only in terms of the aforementioned job attributes. The design details of these two questions are detailed in the section on the online platform.

Sampling frame and timeline Eligible job seekers (to be randomly selected to be invited to participate in the experiment) were selected based on six eligibility criteria: (1) they are registered as unemployed, (2) they are available to start working immediately, (3) they registered their main 5 job search criteria: occupation, reservation wage, commuting distance, type of contract and working hours, (4) they are aged at least 18, (5) they live in Metropolitan France and (6) they have a valid email address and agreed to receive informational emails from the PES.

In order to be representative of the whole job seeker population, the 20 000 job seekers invited to participate in the pilot study were selected through a stratified random sampling from the population of eligible job seekers. The stratifying variables were: age category (4 categories), occupation category (14 categories), and level of assistance received from caseworker (3 categories)³. The pilot study was implemented in two waves, each one targeting around 10 000 job seekers. Each job seeker selected to participate received one invitation email then three reminders, in case of non-response to the survey. For each wave, job seekers were sampled the day before the first invitation email was sent. The first wave was sent on 14 February, 2022 and the second one on 28 March, 2022. In a context of recurrent phishing campaigns, we have taken great care in drafting the invitation email, as this is the first contact with the selected job seekers. The challenge was to explain the purpose of our study and its implications, particularly in terms of the use of their personal data, while encouraging them to participate. The email sent is displayed in Figure D.4 in appendix.

1.2.4 Large-scale experiment: experimental design and sampling frame

Experimental design The large-scale experiment involves three treatment groups of equal size. The main treatment, referred to as the “full treatment” involves incorporating individual preferences into the PES matching algorithm by using personalized weights, as defined by equation 2.5. These individual weights are gathered through the constant sum questionnaire described above. However, this preference declaration task might generate some side effects that could hinder the estimation of the pure impact of recommendation personalization: the preference declaration task may help job seekers to better

³When job seekers register at the French PES, caseworkers assess their level of autonomy and assign them to one of three assistance tracks (strong, medium, light).

identify and prioritize their job search criteria, but it might also make them more demanding regarding the quality and the level of personalization of the recommendations, given that they have made the effort to declare their personal preferences.

To further investigate these potential side effects, we use a “partial treatment” group. In this group, individual preferences are collected, but job recommendations are generated using the standard PES algorithm based on pre-existing weights. A third group, the “control” group, does not participate in the preference declaration task and receives recommendations directly calculated by the standard algorithm. This experimental design enables us to differentiate between the impact of simply expressing preferences and the actual implementation of those preferences within the recommendation algorithm.

The partial treatment group is particularly valuable for understanding whether the act of stating preferences influences the appreciation and adoption of recommendations, independent of any algorithmic adjustments. Importantly, participants were not explicitly informed that their preferences would be used to tailor the recommendations they receive. The recommendations presented to them were simply framed as “job recommendations based on your search criteria”. This careful choice of language, along with the design of our online platform, was intended to prevent participants from developing unrealistic expectations regarding the use of their stated preferences. Instead, the recruitment message emphasized the significance of their contribution to the testing of a new service, thereby aligning their expectations with the broader goals of the experiment. Detailed descriptions of the interface screens used in each treatment condition are provided in the section below.

All job seekers assigned to the three treatment groups were invited with the same email to participate in the test of a new job referral service developed by the PES. Those who volunteered to test the service accessed our online platform via a link in the email. It is only after accessing this online platform that the user experience diverged among the treatment groups: (1) only the full and partial treatment groups were required to declare explicit individual preferences about the relative importance of their job search criteria, and (2) only the full treatment group received algorithmic recommendations personalized according to their preferences, while the partial treatment and control groups received recommendations generated by the standard algorithm. The different experimental conditions according to treatment groups are summarized in table 2.2.

Sampling frame and timeline The eligibility criteria for being randomly assigned to one of the treatment groups are the same as those for the pilot study. Eligible job seekers were assigned to the three treatment groups via stratified randomization according to the pre-treatment volume of job postings published on the PES job board within the

	Email invitation	Collecting preferences	Recommendations are generated by
Full treatment	Yes	Yes	Personalized algorithm
Partial treatment	Yes	Yes	Standard algorithm
Control	Yes	No	Standard algorithm

Table 2.2: Treatment groups

job seekers’ relevant micro-markets. Job seekers’ relevant micro-markets are defined as the intersection between their commuting zone and the occupation they target in their job search criteria. The commuting zones framework we use to define micro-markets is the “Bassins d’emploi” database, which is administered by the PES. It is composed of 404 commuting zones, defined as homogeneous geographical areas where most of the workforce resides and works, and where firms can find most of the labor needed to fill the jobs offered. Job seekers are mapped to a specific commuting zone through their residential location. The occupation database we use to define micro-markets is the “ROME” database, also administered by the PES and which includes 11 097 distinct occupations. Each occupation has an average of 137 “adjacent” occupations, defined as the jobs that a person could also perform, i.e. without additional training or experience. For each job seeker, we computed the number of online job vacancies in their relevant micro-market (commuting zone \times occupation sought enlarged by its adjacent jobs) on April 26, 2022, i.e. approximately two weeks before randomization. This quantitative variable was then divided into 5 categories that were used for stratification (see table 2.3 to get an idea of the distribution of the variable). The stratified randomization was done the day before the emails were sent.

The email campaign started on May 10, 2022 and consisted in 1 invitation email then 3 reminders for job seekers who did not participate. Each individual invited received the same email, regardless of treatment group. In a context of recurrent phishing campaigns, we have taken great care in drafting the invitation email, as this is the first contact with the selected job seekers. The challenge was to explain the purpose of our study and its implications, particularly in terms of the use of their personal data, while encouraging them to participate. The email sent is displayed in Figure D.5 in appendix.

1.3 The online platform

We designed an online platform to host the two stages of our experiment: the pilot study and the large-scale experiment. The interface has three features: (1) it is designed to collect job seekers’ preferences regarding the relative importance they give to job attributes

	Number of eligible job seekers	Share
Less than 4	340 340	20.2%
From 5 to 19	510 585	30.3%
From 20 to 49	418 211	24.8%
From 50 to 99	225 963	13.4%
More than 100	187 945	11.2%

Table 2.3: Number of vacancies online as of April 26, 2022 (before randomization)

(occupation, reservation wage, commuting distance, working hours and type of contract) which are then turned into weights for use in the algorithm; (2) it allows to check whether the preferences collected in the constant-sum question (feature 1) correspond to the implicit preferences of the job seekers; (3) it recommends vacancies (using either the standard algorithm or the personalized algorithm according to the person’s weights) on which job seekers can click to view their full details and apply if they wish. The interface was beta-tested during a focus group with job seekers⁴. Finally, the interface was developed by the French PES.

Once the job seeker clicks on the link in the invitation email and gets to the interface, their user journey proceeds as represented in the Figures C.2 and C.3 in the appendix and as described below.

1.3.1 Homepage

The homepage introduces job seekers to the purpose of the experiment, the nature of the data collected, and the anonymization process employed. Participants are required to log in using their Pôle emploi credentials, which serves two purposes: (1) accessing the values of their main job search criteria, and (2) obtaining their consent to participate in the experiment. By logging in, job seekers agree to take part in the study; those who do not wish to participate can simply leave the webpage.

⁴The focus group took place in Lille, France in July 2019. Eight job seekers were initially invited to participate; seven of them actually showed up and attended the 2 hours and a half discussion. Job seekers were selected according to their personal characteristics, in order to have a sample of representative job seekers. The clarity and feasibility of the whole survey was assessed during this session. Beta testers well understood the constant-sum question, as well as the ranking exercise used in the pilot study.

1.3.2 First screen: measuring preferences that can be used as weights

The first screen (see Figure D.7 in appendix) is designed to capture job seekers' preferences using a constant-sum question. Participants are allocated 15 points to distribute among five job attributes: occupation, commuting distance, wage, contract type, and working hours. The choice of 15 points allows for a balanced allocation across attributes, while still enabling participants to express varying levels of importance. The attributes are tailored to each job seeker's specific search criteria, enhancing relevance and reducing ambiguity. To prevent bias from the order of presentation, the attributes appear in a random sequence. Participants must allocate all 15 points before proceeding to the next screen.

1.3.3 Second screen (only for pilot study): assessing the reliability of the preference measurement

On the second screen (see Figure D.8 in appendix), participants are asked to rank eight job vacancies based on the five key criteria collected earlier —occupation, location, salary, contract type, and working hours. These vacancies are selected to ensure sufficient variation, allowing for meaningful rankings⁵. These vacancies are shown the job seeker in a random order. Participants rank the vacancies using a drop-down menu, with the option to rank as few as six out of the eight vacancies if they find it difficult to distinguish among them. This task is included to verify the consistency of preferences expressed in the previous step. In rare cases where the algorithm does not return any vacancies, an error message is displayed, and the participant cannot proceed.

We deliberately placed the ranking task after the weighting task to optimize the user experience and ensure that participants were not discouraged by the complexity of the tasks. The weighting task, which involves allocating points to different job attributes, is conceptually simpler and more intuitive. By starting with this easier task, we aimed to build the participants' confidence and provide them with a clear understanding of the job attributes they would later use to rank the ads. This sequence helps ensure that when participants proceed to the more complex ranking task, they already have a well-formed idea of the relative importance of each attribute, making the ranking process more straightforward and aligned with their preferences.

⁵In order to make the ranking easier for the job seekers, selected vacancies are sufficiently different from each other: we have not selected the top-8 but the top-1, the top-(1+k), the top-(1+2k), the top-(1+3k),... and the top-(1+7k), where $k = \min(\lfloor \frac{L-1}{7} \rfloor; 10)$, with L the length of the list returned by the algorithm.

1.3.4 Third screen: recommending vacancies

The third and last screen (see Figure D.9 in appendix) is designed to show the participant the whole list of job ads recommendations, either from the PES matching algorithm with individual weights (collected through the first screen) or from standard weights, depending on treatment group. If respondents click on a job ad, they are redirected to the detailed vacancy available on PES job portal.

Stated preferences were intentionally not incorporated into the recommendation algorithm for job seekers in the partial treatment group. However, the language used to present these recommendations—“Here are job recommendations based on your search criteria”—was carefully crafted to be neutral, ensuring that it did not imply that the recommendations were specifically tailored based on the preferences they had provided. The interface design, as detailed earlier in this section, was intentionally structured to guide participants through the process without fostering an expectation that their individual preferences would directly influence the recommendations they received. Moreover, the communication throughout the experiment consistently emphasized that the purpose of collecting preferences was to aid in improving the service for future users, rather than to deliver immediate, personalized recommendations. This approach was aimed at ensuring that participants did not feel misled, but rather understood that their input was being used for broader, long-term improvements.

1.4 Technical challenges and unanticipated issues

Implementation error: sorting mechanism misconfiguration An unforeseen issue arose due to an error in the development of the online platform. While the interface was designed to sort vacancies by relevance, reflecting the personalized preferences of each job seeker, a misconfiguration led to the ads being sorted by their creation date instead. Consequently, all participants, regardless of their treatment group, were shown job ads ordered by the date they were posted rather than by how well they matched individual preferences.

This error was not immediately apparent and only came to light when we scrutinized the initial results. Upon further investigation, the log data revealed that the sorting parameter had been incorrectly set to “date” instead of “relevance”. This error, which was beyond our control and could not have been anticipated by our team, was only discovered after the experiment had concluded and we had collected all the data.

The implications of this error are significant, as it undermined the core objective of our intervention, which was to personalize job recommendations based on individual preferences. The incorrect sorting diminished the impact of this personalization, as the

difference between the recommendations generated by the personalized algorithm and those generated by the standard algorithm was less pronounced. This led to a substantial overlap between the rankings generated by the two algorithms, diluting the observable impact of the personalized recommendations. These consequences are further explored in Section 5.2, where we discuss how this error affected the treatment intensity and reduced the differences between the treatment groups.

Despite this issue, it is important to note that all treatment groups were uniformly affected, allowing the experiment to remain valid as a field study. The error was limited to the display of job recommendations and did not affect the underlying data or the personalized recommendations themselves. Consequently, while the error reduced the intensity of our intervention, the experiment still provides valuable insights into the behavior of job seekers and the relative effectiveness of personalized versus standard recommendations.

Non compliance due to email redirection bug In addition to the sorting mechanism issue, another significant problem arose due to an unexpected bug affecting the emails sent to participants in the large-scale experiment. Like the previous issue, this bug was entirely beyond our control and was discovered only when we began analyzing the results.

The bug impacted how some participants accessed the experiment's interface. Specifically, when participants clicked on the link provided in their invitation email, some of them were redirected to the control group rather than their assigned treatment group. The URL structure for the different groups was as follows: the full treatment group was directed to experimentation-crest.pole-emploi.fr/F5K, the partial treatment group to experimentation-crest.pole-emploi.fr/MX9, and the control group to experimentation-crest.pole-emploi.fr/2DL.

The root cause of this issue likely stems from security features implemented by certain email providers or webmail applications. These providers sometimes rewrite URLs to include tracking parameters or to redirect through a proxy server for security reasons, such as to scan the link for phishing threats or to anonymize the original URL. This process can result in additional parameters being appended to the URL. However, our system was configured to redirect any URL with additional parameters to the control group as a safeguard, which inadvertently caused participants meant for the treatment groups to be rerouted incorrectly.

It is important to emphasize that this error does not impact the internal validity of the experiment. The bug was not influenced by the randomized treatment allocation. As a result, while it affected the distribution of participants across treatment groups, it did so in a manner that was independent of the intended experimental design. This ensures that

the overall integrity of the experiment remains intact. Further exploration of this issue and its impact on the data is discussed in greater detail in Section 2.2.

2 Data and samples statistics

2.1 Data

Our study relies on different types of data, some collected by the PES and others collected directly during our experiment, that can be matched at the individual level.

Pre-treatment data We utilize two comprehensive datasets derived from the rich administrative records collected by the Public Employment Service (PES), one focusing on job seekers and the other on job advertisements.

The first dataset provides detailed information about job seekers, encompassing sociodemographic characteristics such as age, gender, geographical location, marital status, and number of children. Additionally, it captures the job seekers' specific search criteria, including the occupations they are targeting, their preferred type of contract (permanent or fixed-term), the maximum acceptable commuting distance, their desired working hours (full-time or part-time), and their reservation wage. This dataset also includes extensive resume information, reflecting the job seekers' educational background, work experience, language skills, and possession of a driving license. Furthermore, it documents the job seekers' unemployment history, detailing the duration of their unemployment, their unemployment benefits status, their level of autonomy in the job search process, and the reasons for their unemployment.

The second dataset focuses on job advertisements, containing detailed attributes of the jobs being offered. This includes information on the occupation, type of contract, geographical location, working hours, and wage associated with each job ad. By combining these two datasets, we are able to conduct a nuanced analysis of the job matching process, examining how the characteristics of job seekers align with the attributes of available job advertisements and how this alignment influences job seekers' engagement with the recommendations provided by the PES platform.

Online platform data The data collected through the online platform during both the pilot study and the large-scale experiment include several key elements. On the first screen, explicit preferences are captured through the allocation of points among job attributes. In the pilot study, implicit preferences are also gathered on the second screen, where job seekers rank a randomized list of vacancies; the original order of this list is

recorded to enable comparisons of reordering behavior. In both the pilot and large-scale experiments, the system logs the list of job ads viewed on the final screen, along with the vacancies that would have been returned by the standard PES matching algorithm. The data finally include information on the matching scores associated with the job ads recommended to each job seeker. For participants in the full and partial treatment groups, personalized matching scores are recorded, reflecting the extent to which the job ads align with the individualized preferences stated by the job seekers. For all participants, including those in the control group, standard matching scores are provided, representing the degree of alignment based on the pre-existing weights used by the PES algorithm. Additionally, the date and time of each connection to the online platform are tracked.

Outcomes in the large-scale experiment Our primary objective is to assess how the personalization treatment affects user engagement and interest in job recommendations. We focus on interactions that occur during the job seekers' first connection to the online platform to isolate the effect of the recommendations themselves, independent of the total number of platform visits, which could artificially inflate engagement metrics. Specifically, we analyze the number of clicks job seekers make on the recommended job listings during this initial session. A higher number of clicks suggests greater engagement and interest. However, click counts alone do not necessarily distinguish between job seekers who find the recommendations relevant and those who click out of curiosity, only to find the jobs unsuitable upon further review. To address this, we also examine deeper measures of engagement by evaluating the number of times job seekers click the "apply" button on a job vacancy page, which signals an intention to apply, as well as the actual number of job applications submitted. From a policy perspective, increasing the number of job applications is particularly significant, as it represents the crucial step from passive job searching to active engagement in the labor market, a key objective for public employment services. In addition to these measures, we consider the number of connections to the online platform as a key outcome variable. This metric reflects job seekers' interest in the platform itself, especially as this recommendation service is not as easily accessible on the PES website. It serves as an important indicator of the perceived value of the service.

2.2 Description of the samples of participants

2.2.1 Pilot study

Participation rate The sample that was invited to participate in the pilot study was composed of 21,063 job seekers. We observed that 77% of them opened at least an invitation email, 24% visited the interface, whereas only 11% logged in and declared their prefer-

ences through points allocation, and 6% ranked the randomized list of job ads. The fact that not everyone completes the two survey tasks is mainly due to job seekers who do not wish to log in to the interface, but also to the fact that some people were excluded from the experiment just after logging in because of an insufficient number of vacancies recommended by the matching algorithm (a number lower than 8 making it impossible to carry out the vacancy ranking exercise). Unfortunately, our data do not enable to disentangle between the two effects. Ultimately, 2.234 job seekers gave their explicit preferences but only 1.740 job seekers gave us insight on their implicit preferences through job ads reordering.

Definition of the subsample for analysis The sample of respondents (who completed the weight allocation task) is not representative of the population of job seekers registered at the PES (see table E.1 in the appendix). Demographic variables indicate that participants to the experiment are on average 44 years old (compared to 38 years old in the full sample), and that about 56.6% are female (52% in the full sample), 49% are married (40% in the full sample), 40% have a university degree (33% in the full sample) and 63% are unemployed for the first time (54% in the full sample). Highly skilled job seekers are also over-represented among participants (23% vs. 15% in the full sample) and the average participant has a gross monthly reservation wage of about 2032 euros, compared to 1897 euros in the full sample.

Women are more likely to participate to the pilot study than men (56% of respondents are women, against 51% for the full sample), respondents are also older than non-respondents (they are on average 6 years older than non-respondents). Consistently, respondents have on average 40 months longer work experience than non-respondents. Respondents are slightly more educated than non-respondents: 29% of respondents have a college education level whereas only 24% of non-respondents have a college education level. The conclusion is even more striking when dealing with skills of job seekers: 23% of respondents are looking for high-skilled occupations, whereas only 14% of non-respondents are.

In addition to sociodemographic characteristics, it appears that there are also disparities in job searches of respondent and non-respondent job seekers. Consistent with what was described above, respondent job seekers have a reservation wage on average 9% higher than non respondent job seekers' reservation wage (respectively 2 058 € and 1 888 €). We also find the salary-distance trade-off often discussed in the literature insofar as, on average, respondents declare a significantly higher maximum commuting time than non-respondents (more than an extra minute on average, corresponding to almost +5%). In line with the more precarious situation of non-respondents job seekers on average,

they search for temporary contract and part time job more often than respondent. Finally, the sectors of activity sought by job seekers vary noticeably between respondents and non-respondents to our survey: searches in the business support services are particularly over-represented among respondents (25% of the searches compared to 19% of the non-respondent job seekers' searches), while Construction and public works, Tourism and leisure, and Transport sectors are significantly more present among non-respondent job seekers' searches (representing 28% of the searches, against 21% of respondent job seekers' searches).

2.2.2 Large-scale experiment

Randomization check Table E.2 in the appendix shows descriptive statistics on pre-intervention variables by treatment group, calculated on the full sample that was invited to participate in the experiment. Only one out of the 43 balancing tests is significant (at the 5% level), which confirms that the randomization was successful at balancing the treatment groups.

Participation rate Regarding participation in our experiment, we observed that 81% of email recipients opened at least one invitation email (among the 4 emails of our campaign), but only 12.6% of them logged in at least once to the interface to participate in the experiment. The difference between the email read rate and the participation rate can be partly explained by several factors. Some job seekers may not have been able or willing to log in to the interface due to reasons such as lost login credentials, concerns about the confidentiality of their personal data, or a reluctance to participate that emerged after reading the text on the home page of the interface. In the following, we refer to individuals who logged into the platform as the *participants to the experiment*, a key distinction for the remainder of our analysis.

Definition of the subsample for analysis Among the sample of participants, we observed that approximately 85% of participants assigned to the full and partial treatment groups did comply with their assigned treatment. This non-compliance is attributed to two primary factors: the bug that unintentionally redirected some participants to the control group as well as voluntary attrition where participants chose to exit the interface without completing the point allocation. Table 2.4 provides a detailed breakdown of the treatment assignment versus the treatment received. Notably, 84.98% of participants assigned to the full treatment group received the treatment as intended (completed the weight entry task and saw recommendations), while 5.83% were mistakenly redirected to

the control group, and 9.19% exited the interface without completing the weight allocation. Similarly, for those assigned to the partial treatment group, 84.63% completed the treatment as intended, 5.85% were redirected to the control group due to the bug, and 9.52% exited without completing the weight allocation. The control group, by design, did not experience any reassignment or attrition related to the weight allocation task, with 100% of participants remaining within the control condition as expected.

Table 2.4: Assigned treatment vs. received treatment for participants who logged in

	Logged into full treatment	Logged into partial treatment	Logged into control	Logged but did not complete task
Assigned to full treatment	84.98%	0.00%	5.83%	9.19%
Assigned to partial treatment	0.00%	84.63%	5.85%	9.52%
Assigned to control	0.00%	0.00%	100%	0.00%

Note: “Logged in but did not complete the task” refers to participants who logged into their assigned treatment group but exited the experiment without completing the task. These individuals did not proceed to control or any other group but simply did not finish the task.

Importantly, Table 3.1 shows that treatment assignment did not affect participation to the experiment. This was expected given our experimental design : all the job seekers invited to participate have received exactly the same invitation email. Therefore, in the following, we (safely) restrict our analysis sample to those who logged in at least once. Table 3.1 additionally shows that treatment assignment into full or partial treatments did not affect the exit rate (defined as not completing the weight question) or non-compliance (being redirected towards the control group). This was also expected: the exit of job seekers occurs during the weight completion exercise and at this stage, full and partial treatments still have the same user experience ; the bug affected individuals from full and partial treatments before they even logged into the platform.

We finally investigate the representativeness of the participant sample along the same dimensions as in table E.2. Table B.3 in the appendix shows summary statistics on the pre-intervention characteristics of the full sample that was invited to participate compared to the sample of participants (those who logged in at least once to the interface). The participant sample is composed of 31.438 job seekers. Demographic variables indicate that participants to the experiment are on average 44 years old (compared to 38 years old in the full sample), and that about 56.6% are female (52% in the full sample), 49% are married (40% in the full sample), 40% have a university degree (33% in the full sample) and 63% are unemployed for the first time (54% in the full sample). Highly skilled job

Table 2.5: Treatment differences in participation and issues

	Full sample	Sample of participants (who logged in) <i>among full & partial treat. groups</i>	
	Logged in to the interface (1)	Not affected by the bug (2)	Completed the weight entry task (3)
Assigned to partial treatment	0.002 (0.002)	-0.0002 (0.003)	-0.003 (0.004)
Assigned to control	0.001 (0.002)		
Mean value full treat. group	0.125	0.942	0.905
N	249 918	20 845	20 845

Note: The table reports treatment differences regarding : (1) having logged in at least once to the interface ; (2) not being affected by the bug that redirected job seekers assigned to the full and partial groups to the control group ; (3) having completed the weight allocation question when assigned to the full and partial groups. Model (1) is estimated on the full sample of job seekers that received an email invitation to participate in the experiment, models (2) and (3) are estimated on the sample of participants (i.e they logged in at least once to the interface) which were assigned to full and partial treatment groups. Robust standard errors are reported in parentheses. *, **, ***: significance at 10%, 5% and 1%.

seekers are also over-represented among participants (23% vs. 15% in the full sample) and the average participant has a gross monthly reservation wage of about 2032 euros, compared to 1898 euros in the full sample.

3 Results from the pilot study: evaluating the reliability of the preference measurement

3.1 Respondents' engagement with the survey tasks

In this section, we analyze job seekers' engagement with the two tasks that comprised the pilot survey: the point allocation question and the ranking exercise. Despite a lower completion rate for the ranking task, the data suggests that job seekers engaged thoughtfully with both activities. On average, respondents spent 2.24 minutes on the point allocation question and 3.72 minutes on the ranking exercise (see Figure F.10 for the distribution of time spent). The additional time required for the ranking task reflects the greater cognitive effort involved in reordering job ads compared to distributing weights across attributes. Notably, there is no evidence of rushed behavior, as only 1.6% of job seekers completed the ranking task in less than 1 minute, a time frame too short to reflect meaningful engagement.

To further assess the quality of engagement, we examined two indicators: straightlining (assigning identical responses across all options) and the modification of the initial random ranking presented to job seekers. In the point allocation task, only 0.5% of job seekers assigned equal weights to all criteria, indicating a thoughtful approach. Similarly, only 2.4% of respondents left the initial random ranking of job ads unchanged, suggesting that most participants took the time to reorder the ads according to their preferences. Moreover, just 1.9% of job seekers were shown a random ranking that coincidentally matched the personalized recommendation from the PES algorithm.

These findings indicate that, while some job seekers may have found the ranking task more cognitively demanding, those who completed it invested time and effort. The low rates of straightlining and unaltered rankings further suggest that most participants approached the tasks with genuine engagement and attention.

3.2 Are preferences revealed through the ranking consistent with those obtained through the point allocation question?

In this section, we assess the degree of consistency between job seekers' preferences elicited through the point allocation question and those revealed by the ranking task. The objective is to determine whether the personalized matching algorithm, which is based on the job seekers' own stated weights, can replicate their ranking of vacancies. In an ideal scenario, if job seekers' preferences are perfectly captured by the point allocation question, their ranking of vacancies would match the personalized ranking generated by the algorithm. However, this exact match is unlikely due to the complexity of ranking eight distinct job ads, making perfect consistency rare. Therefore, instead of expecting a perfect match, we aim to evaluate whether job seekers' rankings resemble the personalized ranking more closely than a randomly generated ranking of vacancies. This would indicate that job seekers' preferences, as revealed through the ranking task, align more with their personalized preferences than with random orderings.

To quantitatively evaluate the level of convergence between the different rankings, we use the Discounted Cumulative Gain (DCG) metric (Järvelin and Kekäläinen [2002]), which is widely employed in information retrieval. The DCG measures the effectiveness of a ranking system by accounting for both the relevance of ranked items and their positions in the list. Relevance scores are typically based on human judgments or inferred from user interactions (such as clicks), and these scores are transformed using a logarithmic function to discount the importance of lower-ranked items. This reflects the intuition that users are more interested in items at the top of the list. The cumulative gain is then computed across ranking positions to yield the DCG score, with a higher score indicating

a more relevant ranking.

In our context, the relevance of each job vacancy can be measured by the matching score from either equation 2.1 or equation 2.5, depending on the weights used. For a job seeker i who is recommended a list of p vacancies, each with a matching score s , the DCG is computed as:

$$DCG_i^s(R) = \sum_{j=1}^p \frac{s(i, j)}{\log_2(R_{ij} + 1)} \quad (2.8)$$

where R_{ij} represents the position of vacancy j in the ranking. Higher DCG values correspond to rankings where more suitable jobs appear near the top, as the weighting $\frac{1}{\log_2(R_{ij}+1)}$ ⁶ penalizes items ranked lower in the list.

DCG is particularly well-suited for this analysis because it takes into account the matching score associated with each job ad, avoiding heavy penalties when job ads with similar scores are swapped in the ranking. This is crucial in cases where ads have nearly identical scores, as penalizing such inversions would not accurately reflect the job seeker's true preferences.

We compute the DCG for three different rankings of job ads: (1) the ranking generated by the PES algorithm using personalized weights (which we refer to as the "ideal DCG" or $iDCG$, as it ranks the most relevant vacancies at the top by definition); (2) the job seeker's own ranking, denoted $DCG_{jobseeker}$; and (3) the random initial ranking shown to the job seeker during the ranking task, denoted DCG_{random} . We are particularly interested in the following comparisons:

$$\Delta DCG_{ideal,random} = iDCG - DCG_{random} \quad (2.9)$$

$$\Delta DCG_{ideal,jobseeker} = iDCG - DCG_{jobseeker} \quad (2.10)$$

$$\Delta DCG_{random,jobseeker} = DCG_{random} - DCG_{jobseeker} \quad (2.11)$$

Perfect consistency would require $\Delta DCG_{ideal,jobseeker} = 0$, indicating that the job seeker's ranking perfectly matches the personalized algorithm's output. However, a more realistic expectation is that job seekers should at least perform better than a random ordering of vacancies. To assess this, we conduct two checks. First, we evaluate whether job seekers' rankings outperform the random ranking by checking if $\Delta DCG_{random,jobseeker} \leq 0$. This condition tests whether the job seeker's ranking yields a DCG score that is at least as high

⁶A graph illustrating the inverse of the function $\log_2(1+x)$ for $x \in [1, 8]$ can be found in appendix B of this paper

as the random one. If this holds, it means that the job seeker organized the vacancies in a way that better reflects their preferences than random ordering would. Second, we examine how closely the job seeker’s ranking aligns with the ideal algorithmic ranking by assessing whether $\Delta DCG_{ideal,jobseeker} \leq \Delta DCG_{ideal,random}$. This comparison evaluates whether the job seeker’s ranking is more consistent with the personalized algorithm than with the random ranking. In rare cases where the random ranking happens to be close to the ideal, this check may not hold, but such occurrences are exceptions.

Importantly, our results show that 59% of job seekers perform better than random, demonstrating a meaningful level of alignment between job seekers’ preferences and their rankings. Although perfect consistency is rare (only 2.5% of job seekers’ rankings perfectly match the personalized algorithm), the majority of job seekers show some degree of convergence. Furthermore, as illustrated in Figure 2.1, the distributions of deviations from the ideal DCG highlight that job seekers’ rankings (in red) generally deviate less from the ideal ranking than the random rankings (in yellow). This suggests that job seekers tend to rank job ads in a way that is closer to their personalized algorithm’s output than a randomly ordered list. In other words, job seekers’ rankings are closer to the personalized algorithm’s output than to a randomly ordered list of job ads.



Figure 2.1: Distributions of deviations from ideal DCG

To further explore the relationship between these tasks, we examine how the likelihood of performing better than random depends on the degree of “potential gain” from

reordering the random ranking. The potential gain reflects how much a job seeker could improve the ranking by rearranging the initially random list of vacancies:

$$\text{Potential gain} = \frac{iDCG - DCG_{\text{random}}}{DCG_{\text{random}}} \quad (2.12)$$

If the random ranking is already close to the ideal, the potential gain is small, and job seekers have less room to outperform it. In contrast, when the random ranking deviates significantly from the ideal, the potential gain is large, and job seekers are more likely to improve. This variation in potential gain across individuals might explain why some job seekers are better able to perform above random: they start with a random ordering that offers greater room for improvement.

As shown in Figure 2.2, the probability of doing better than random increases with the potential gain. When the random ranking is already close to the ideal ranking, the potential gain is small, and the likelihood of improving on the random ranking is lower.

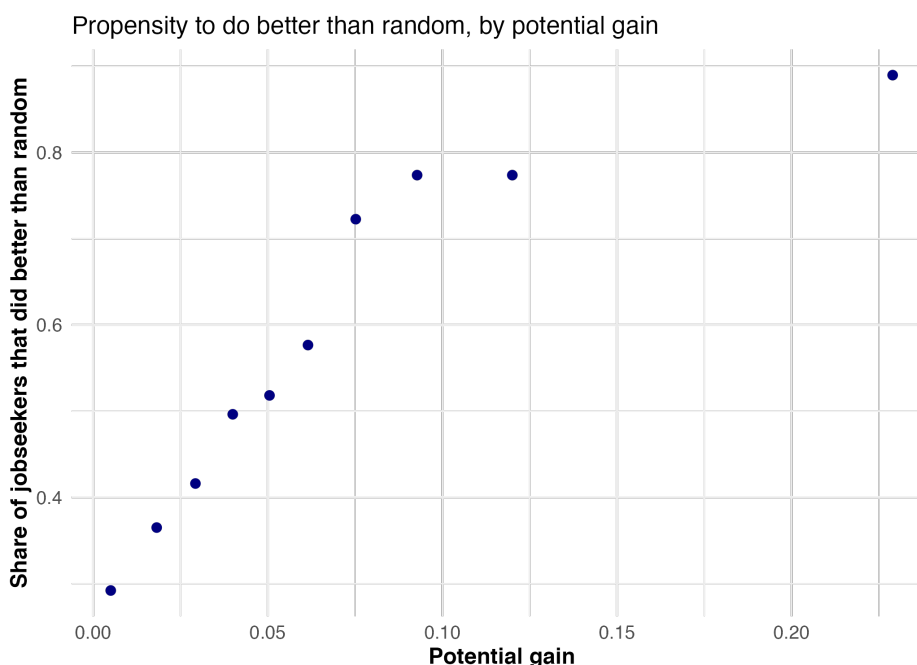


Figure 2.2: Propensity to do better than random, according to potential gain

These results suggest that job seekers are generally able to perform better than random when ranking job ads. Although there is high heterogeneity in the degree of consistency, our findings indicate a moderate level of convergence between preferences collected from the point allocation question and those revealed through the ranking task.

3.3 Which weights fit better revealed preferences: standard or personalized?

In this section, we treat the ranking of job vacancies provided by job seekers as a way to reveal their true preferences. Using this revealed preference data, we assess which set of weights—personalized or standard—better captures these preferences. This approach serves as an alternative method to validate the reliability of the weighting task.

We compare two sets of rankings: those generated by the personalized algorithm, which uses the job seekers' own stated weights, and those produced by the standard algorithm, which applies uniform weights across all attributes. By treating the rankings provided by job seekers as an indicator of their actual preferences, we examine which set of weights—personalized or standard—produces rankings that align more closely with these preferences. Ideally, if the personalized weights accurately capture job seekers' preferences, the personalized ranking should align more closely with the rankings provided by the job seekers themselves. Conversely, the standard ranking would reflect a one-size-fits-all approach, which is less likely to match the individual preferences stated by job seekers.

We compare the DCG obtained with personalized weights to the DCG obtained with standard weights using the following equation:

$$\Delta DCG_{personalized,standard} = DCG_{jobseeker, individual weights} - DCG_{jobseeker, default weights} \quad (2.13)$$

Our results show that 70% of job seekers have a greater gain with personalized weights compared to standard weights ($\Delta DCG_{personalized,standard} \geq 0$). This suggests that, in general, personalized weights fit job seekers' revealed preferences better than standard weights. The distribution of the difference between the gain with individualized weights and standard weights is shown in Figure 2.3.

One way to further improve the fit with job seekers' preferences would be to adapt the hyper-parameters involved in calculating the adequacy scores. For example, the algorithm assumes that the wage adequacy score is equal to 0% when the vacancy wage is below 0.89 times the job seeker's reservation wage, but this may not be consistent with job seekers' actual preferences.

Overall, these results suggest that using personalized weights based on job seekers' preferences can improve the relevance of the job recommender system.



Figure 2.3: DCG of jobseeker’s ranking: personalized weights. vs. standard weights

4 Analysis of stated preferences for job attributes

4.1 Average preferences for job attributes

In this section, we examine the job attribute preferences obtained from the large-scale experiment and the point allocation question. A significant discrepancy is observed between the stated preferences for job attributes (i.e., the weights collected during our experiment) and the ad hoc weights utilized in the PES matching algorithm, as illustrated in Figure 2.4. The occupation attribute displays the most notable difference: the default weight in the PES matching algorithm is considerably higher than other weights, at $\frac{10.7}{15}$, while the sample mean is approximately $\frac{2.68}{15}$. The job attributes of the highest importance, on average, can be ordered as follows (with average weights in parentheses):

1. Wage (3.41),
2. Distance (3.36),
3. Working hours (2.9),
4. Occupation (2.68), and

5. Type of contract (2.63).

Figure 2.4 demonstrates that the weight distributions are considerably balanced in comparison to the standard PES weights. This observation is further supported by Figure G.11, which depicts the distribution of Euclidean distances between each individual's weights and the weight pattern comprising equal weights allocated to each attribute (5 points each). Additionally, the figure exhibits the Euclidean distance between equal weights and the default weights (blue dotted line), as well as the distance between equal weights and the pattern that allocates all 15 points to a single attribute (pink dotted line). The majority of the distribution is concentrated between 0 and the vertical bar representing the default weights, signifying that a vast majority of job seekers report weights with greater balance than the default PES weights. Lastly, we observe that a negligible proportion of job seekers either assign all the weight to a single criterion (1.7%) or declare perfectly balanced weights (0.5%).

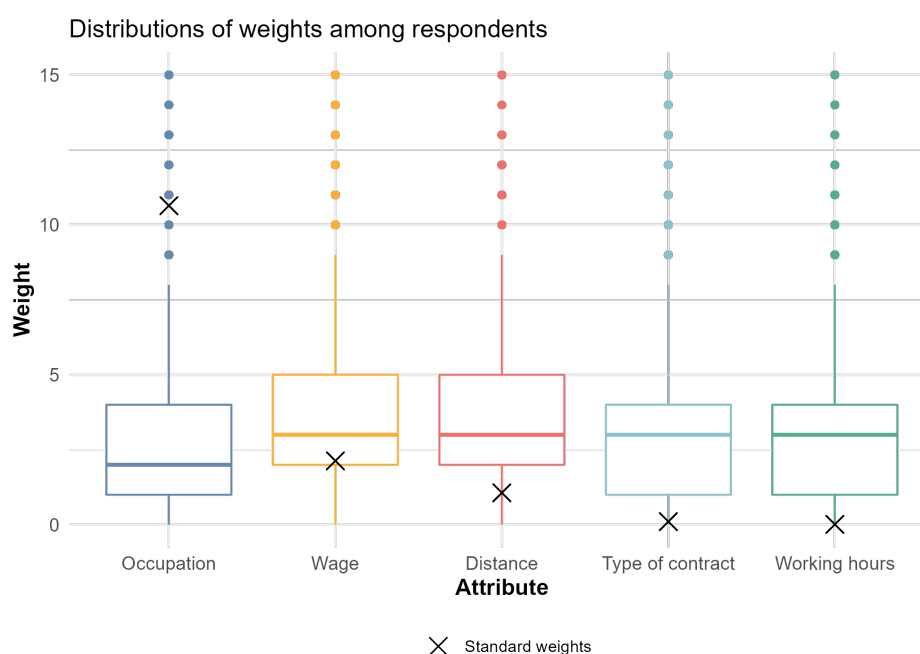


Figure 2.4: Distributions of weights collected and ad-hoc weights used in the PES matching algorithm

4.2 Typical preferences profiles

To further analyze the collected preference data and examine whether the average weights hide some heterogeneity among job seekers, we employed a data-driven approach to reveal typical preference profiles within the population. Using a clustering algorithm

(K-means) on the weights obtained during the large-scale experiment, we derived four distinct stated preference profiles. The clustering only incorporated the weight data collected from the point allocation question, ensuring intra-cluster similarity only in terms of job attributes preferences.

We observe that these four clusters are significantly distinguished by the average weighting profiles they encompass (refer to Figure 2.5). The first cluster (accounting for 9% of the total sample) is the most unbalanced, with job seekers placing substantial importance on the distance attribute relative to the entire sample of respondents (average weight of 8.18). The second cluster (47% of the total sample) represents the most balanced profile, prioritizing the working hours attribute (average weight of 3.94) followed by the type of contract (average weight of 3.83). The third cluster (23% of the total sample) exhibits preferences skewed towards the occupation attribute (average weight of 5.85) more than other attributes. The fourth and final cluster (21% of the total sample) assigns greater importance to the wage attribute (average weight of 6.22).

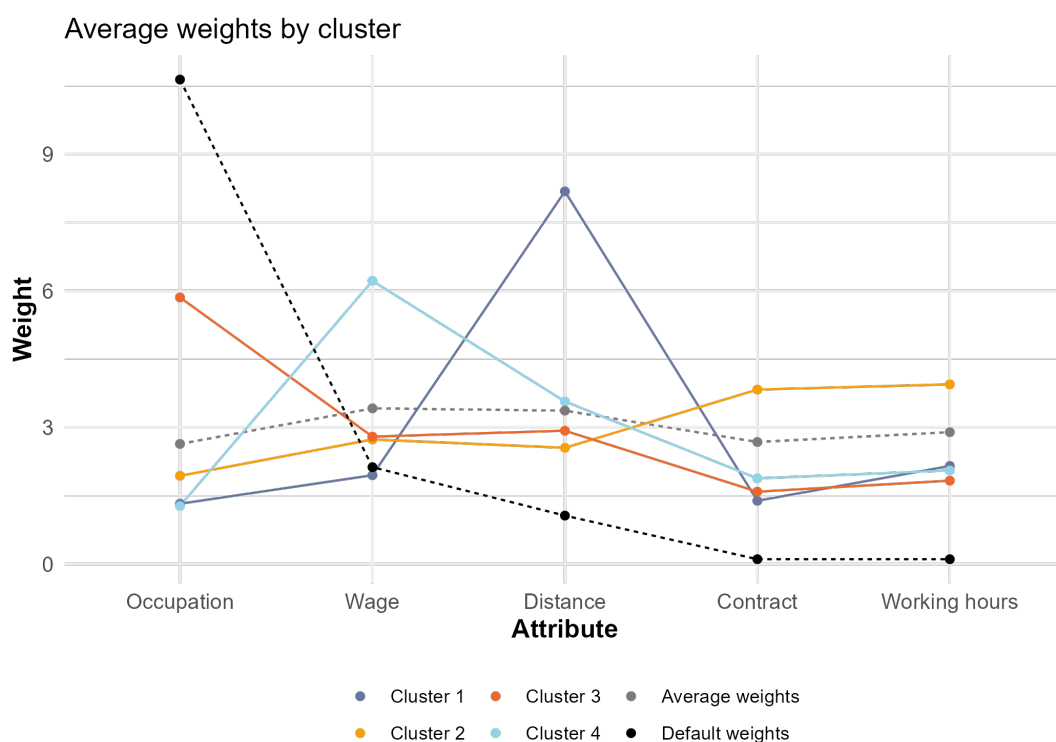


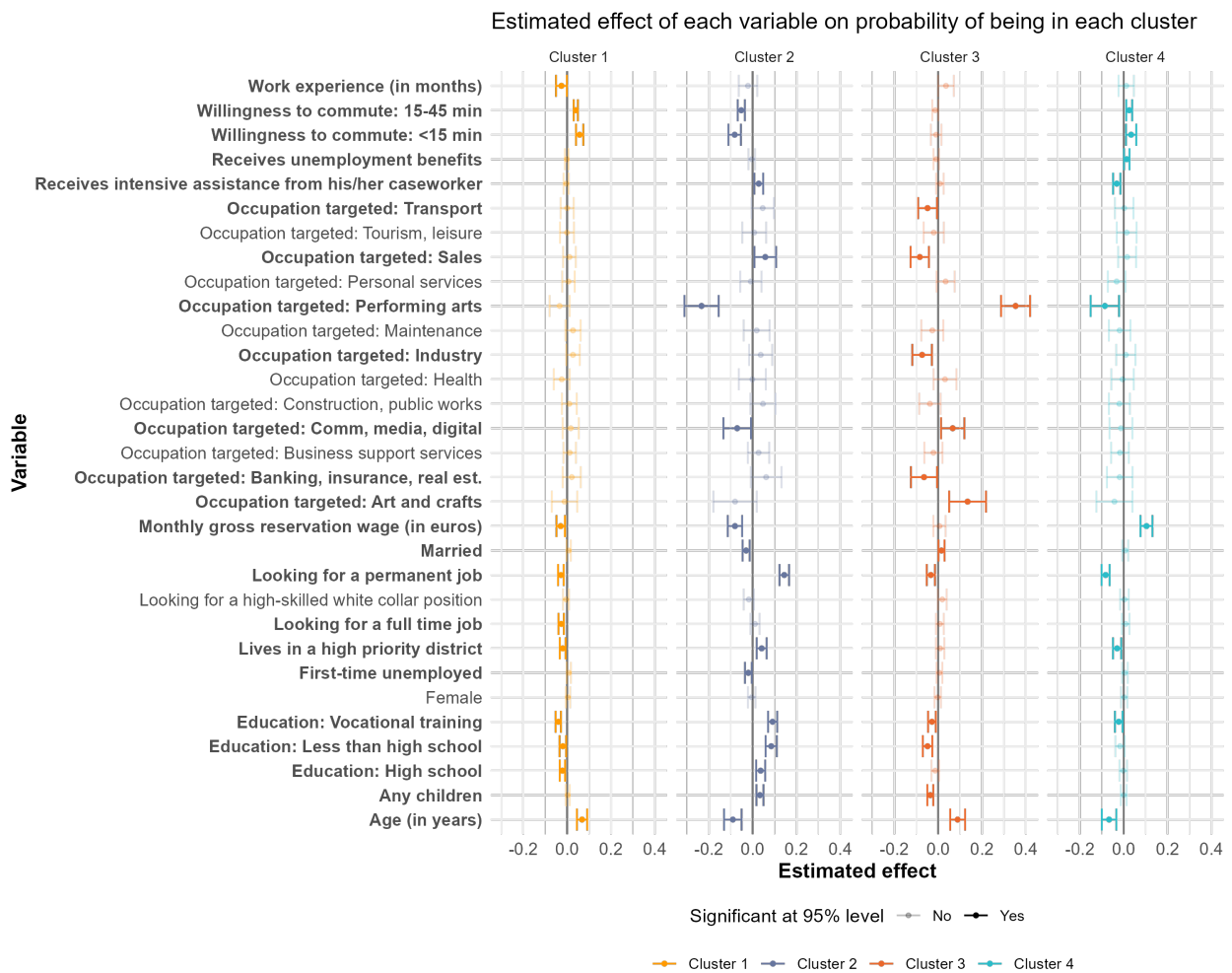
Figure 2.5: Average weights by cluster, compared to average weights and default weights

We proceed to extensively document the characteristics of each preference profile. To interpret the clusters, we initially provide summary statistics on sociodemographic variables and job search criteria for each cluster in Table G.4. Adopting a more data-driven

approach, we further present the results of four linear probability models in Figure 2.6, each aiming to predict an individual's cluster membership based solely on sociodemographic and job search attribute variables. This method enables us to identify which specific covariates have the largest and most significant effects on the probability of belonging to each cluster.

We first note that the job search criteria of job seekers in each cluster align with their preferences concerning job attributes. Job seekers in the first cluster, who highly value the distance attribute, are more likely to have shorter maximum acceptable commutes, seek part-time employment and temporary jobs, and request lower reservation wages compared to the full sample. The second cluster's job seekers, who prioritize working hours and contract type, are more likely to search for a permanent job. Those in the third cluster, valuing more the occupation attribute, generally possess greater experience and are more likely to seek high-skilled and well-paid positions. Lastly, job seekers in the fourth cluster, valuing the wage attribute, demand higher reservation wages.

We now examine the clusters in terms of sociodemographic variables, finding notable differences among the clusters. For the first cluster, job seekers are predominantly female and married. The linear probability model predicting membership in cluster 1 reveals that a shorter maximum acceptable commute time has a substantial and significant effect on the likelihood of belonging to this cluster. Additionally, age exerts a large positive and significant influence on cluster 1 membership, while work experience has a negative and significant impact. The second cluster comprises job seekers with lower education and experience levels, who on average demand lower reservation wages. The linear probability model for this cluster indicates that age and reservation wage have large, negative, and significant effects on the probability of belonging to cluster 2. Seeking a permanent job and having a lower educational background are found to positively and significantly influence the likelihood of being part of this cluster. Job seekers in the third cluster are more likely to possess a college diploma, have fewer children, and exhibit greater work experience and higher reservation wages than others. The linear probability model for this cluster demonstrates that age and seeking specific occupations in the Arts domain have large, positive and significant effects on the probability of belonging to cluster 3. Lastly, job seekers in the fourth cluster display higher experience levels and reservation wages compared to others. The key difference between the third and fourth clusters, aside from their job attribute preferences, lies in the factors influencing the probability of cluster membership. In the fourth cluster, both reservation wage and age have positive, large, and significant impacts on the probability of membership, while in the third cluster, age has a negative and significant effect, and reservation wage does not significantly influence the likelihood of belonging to the cluster.



Note: Each graph reports the estimated coefficients and confidence intervals from an OLS regression trying to predict if the observation belongs to Cluster i or not. The confidence intervals represented are at the 95% level.

Figure 2.6: Estimated effect of each covariate on each cluster's belonging

5 Descriptive analysis of the recommendations generated by the two matching algorithms

5.1 How does weight personalization affect the recommendations made?

In this section, we explore the empirical differences between recommendations generated by the standard PES matching algorithm and those produced by the PES matching algorithm with personalized weights during the large-scale experiment. Both algorithms recommend only the job ads with the highest matching scores among vacancies with a score of at least 50%, up to a maximum of 150 vacancies. Consequently, substantial dif-

ferences may exist between the standard and personalized outputs: the two algorithms might not recommend the same number of vacancies or the same vacancies, and even if they recommend the same vacancies, the rankings are likely to differ. However, the implementation issue encountered during the experiment reduces the variation between the two rankings, as the recommended vacancies are sorted by date rather than relevance in each ranking, making them more similar than expected. Therefore, in this section, we focus solely on the recommendations made by each algorithm, irrespective of the order in which they were made.

In the following, we compare the standard recommendation set for each individual i (denoted as $\mathcal{R}^s(i)$) with their personalized recommendation set (denoted as $\mathcal{R}^p(i)$). Every respondent from the full and partial treatment groups provided their weights, allowing us to generate two rankings for each individual: a personalized ranking based on the individual's weights and a standard ranking based on the algorithm's default weights. During the experiment, job seekers were only shown one of these two rankings, depending on their treatment group. We introduce the notation $\mathcal{R}_{overlap}(i)$ to represent the set of vacancies included in both recommendation sets:

$$\mathcal{R}_{overlap}(i) = \mathcal{R}^s(i) \cap \mathcal{R}^p(i) \quad (2.14)$$

Moreover, we define $\mathcal{R}_{perso. only}(i)$ as the set of vacancies included in the personalized set but not in the standard set:

$$\mathcal{R}_{perso. only}(i) = \{j \in \mathcal{R}^p(i) \mid j \notin \mathcal{R}^s(i)\} \quad (2.15)$$

We also define $\mathcal{R}_{std. only}(i)$ as the set of vacancies included in the standard set but not in the personalized set:

$$\mathcal{R}_{std. only}(i) = \{j \in \mathcal{R}^s(i) \mid j \notin \mathcal{R}^p(i)\} \quad (2.16)$$

Furthermore, we use $\mathcal{R}_{seen}(i)$ to denote the set of ads viewed by job seeker i during the experiment. Job seekers in the full treatment group were shown a set of ads equal to $\mathcal{R}_{seen} = \mathcal{R}^p = \mathcal{R}_{overlap} \cup \mathcal{R}_{perso. only}$. Job seekers in the partial treatment group were shown a set of ads equal to $\mathcal{R}_{seen} = \mathcal{R}^s = \mathcal{R}_{overlap} \cup \mathcal{R}_{std. only}$.

On average, the standard algorithm returns approximately 40 vacancies per job seeker ($|\overline{\mathcal{R}^s}| \approx 40$), whereas the personalized algorithm returns approximately 48 vacancies ($|\overline{\mathcal{R}^p}| \approx 48$). This discrepancy is primarily attributed to the personalized algorithm's ability to generate new recommendations while maintaining the same vacancies recommended by the standard algorithm. Indeed, the average size of the overlapping vacancies set is 36 ads ($|\overline{\mathcal{R}_{overlap}}| \approx 36$), the average size of the personalized only vacancies set is 12

Table 2.6: Summary statistics on recommendations sets

	Full treatment ($N = 8759$)		Partial treatment ($N = 8861$)		Balancing statistics	
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Number of job ads recommended by each algorithm						
$ \mathcal{R}_{seen} $	47.717	53.507	39.962	50.161	-7.755	0.782
$ \mathcal{R}^p $	47.717	53.507	47.241	53.563	-0.476	0.807
$ \mathcal{R}^s $	40.255	49.958	39.962	50.161	-0.293	0.754
Number of job ads in each recommendation subset						
$ \mathcal{R}_{overlap} $	35.938	43.562	35.742	43.931	-0.196	0.659
$ \mathcal{R}_{perso. only} $	11.780	20.618	11.499	20.316	-0.281	0.308
$ \mathcal{R}_{std. only} $	4.317	13.301	4.220	12.832	-0.097	0.197
Impact of treatment: distribution of individuals across the possible scenarios						
$\mathbb{1}(\mathcal{R}_{seen} = \emptyset)$	0.085	0.279	0.090	0.286	0.005	0.004
$\mathbb{1}(\mathcal{R}_{overlap} \neq \emptyset, \mathcal{R}_{perso. only} \neq \emptyset, \mathcal{R}_{std. only} = \emptyset)$	0.446	0.497	0.432	0.495	-0.014	0.007
$\mathbb{1}(\mathcal{R}_{overlap} \neq \emptyset, \mathcal{R}_{perso. only} = \emptyset, \mathcal{R}_{std. only} \neq \emptyset)$	0.025	0.155	0.024	0.153	-0.0007	0.002
$\mathbb{1}(\mathcal{R}_{overlap} = \emptyset, \mathcal{R}_{perso. only} \neq \emptyset, \mathcal{R}_{std. only} \neq \emptyset)$	0.004	0.061	0.006	0.074	0.002	0.001
$\mathbb{1}(\mathcal{R}_{overlap} = \emptyset, \mathcal{R}_{perso. only} = \emptyset, \mathcal{R}_{std. only} \neq \emptyset)$	0.005	0.068	0.004	0.066	-0.0003	0.001
$\mathbb{1}(\mathcal{R}_{overlap} \neq \emptyset, \mathcal{R}_{perso. only} \neq \emptyset, \mathcal{R}_{std. only} = \emptyset)$	0.022	0.146	0.022	0.146	-0.0001	0.002
$\mathbb{1}(\mathcal{R}_{overlap} \neq \emptyset, \mathcal{R}_{perso. only} = \emptyset, \mathcal{R}_{std. only} = \emptyset)$	0.168	0.374	0.166	0.372	-0.002	0.006
$\mathbb{1}(\mathcal{R}_{overlap} \neq \emptyset, \mathcal{R}_{perso. only} \neq \emptyset, \mathcal{R}_{std. only} \neq \emptyset)$	0.246	0.431	0.257	0.437	0.010	0.007

($|\overline{\mathcal{R}_{perso. only}}| \approx 12$), and the average size of the standard only set is 4 ($|\overline{\mathcal{R}_{std. only}}| \approx 4$). This finding becomes more evident when analyzing the distribution of job seekers across various possible scenarios: notably, 46% of job seekers had at least one vacancy recommended exclusively by the personalized algorithm and no recommendations exclusively from the standard algorithm. In this case, the weight personalization treatment solely recommended additional vacancies compared to the standard algorithm, without dropping any recommendations from the standard algorithm. Additionally, for 17% of job seekers, the treatment exhibited no impact, as both algorithms recommended identical vacancies with no exclusive recommendations generated by either algorithm. Table 2.6 presents detailed statistics regarding the sizes of recommendation sets according to the two treatment groups. As anticipated, there is no significant difference between the two treatment groups concerning the recommendations made by each of the two algorithms. The only difference between the groups in the table, measured post-treatment, relates to the number of ads seen (which were generated by the personalized algorithm for the full treatment and by the standard algorithm for the partial treatment).

We now show that the personalized algorithm’s propensity to generate new recommendations stems from the fact that it is generally more challenging for vacancies to obtain a score above 50% (i.e., not be filtered out) with standard weights compared to personalized weights. The standard weights are heavily skewed towards occupation, causing even a minor mismatch between the desired and advertised occupation to significantly reduce a vacancy’s score. On average, personalized weights are more balanced

than standard weights, which results in less variance within scores and subsequently less filtering effects. Table 2.7 displays the average scores associated with each attribute of the vacancies for which the weight was altered during the intervention: occupation, wage, distance, type of contract, and weekly hours. A score closer to 1 indicates a greater alignment between the attribute sought by the job seeker and the attribute offered by the recruiter. This table shows that vacancies newly generated through weight personalization attain a lower occupation score than vacancies generated by the standard algorithm, indicating that the newly generated recommendations encompass occupations that are, on average, further from what job seekers desire. Similarly, newly generated recommendations possess a lower wage score than standard recommendations, implying that they are likely to offer lower wages. Conversely, regarding working hours, the newly generated recommendations are more congruent with job seekers' preferences compared to standard recommendations.

Table 2.7: Matching scores of recommendation sets

Characteristic (std. weight)	$\mathcal{R}_{overlap}$ ($N = 15536$)		$\mathcal{R}_{perso. only}$ ($N = 12631$)		$\mathcal{R}_{std. only}$ ($N = 5022$)	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Occupation (10.7)	0.653	0.181	0.449	0.102	0.669	0.221
Wage (2.1)	0.566	0.345	0.488	0.371	0.480	0.408
Distance (1.1)	0.866	0.178	0.869	0.204	0.775	0.309
Working hours (1.1)	0.745	0.360	0.770	0.362	0.711	0.403
Contract (0.1)	0.490	0.286	0.514	0.328	0.449	0.349
Weighted preference score	0.665	0.157	0.642	0.165	0.543	0.233

Note: The number of observations displayed corresponds to the number of job seekers on which statistics were calculated (since only job seekers who had a non-empty set could be used).

In conclusion, our findings indicate that personalizing the weights of the algorithm leads to job seekers receiving a greater number of recommendations compared to the standard algorithm. These job recommendations tend to deviate, on average, from job seekers' preferences in terms of occupation and salary.

5.2 How did the implementation issue affect the intensity of our treatment?

In this section, we examine the influence of the implementation issue (ads ranked by date instead of matching score) on treatment intensity, specifically by assessing the disparity between the rankings derived from the personalized algorithm and the standard

algorithm. We focus on comparing the top-15 of both rankings, as users typically do not explore the entire list of recommendations. Indeed, most job seekers refrain from scrolling past the top-15, even though more relevant vacancies might be present further down the list (merely 30% of job seekers clicked on recommendations beyond the *top-15*, as illustrated in Figure H.12, which can be found in the Appendix).

As each ranking was presented to job seekers sorted by date rather than by matching score, we only investigate the differences in terms of vacancies included in each top-15 without considering their order. We introduce new notations, corresponding to the recommendations sets limited to top-K.

The standard recommendation set until the top K for an individual i is defined as:

$$R_s^{(K)}(i) = \{j \in J_s(i) \mid R_{sij} \leq K\}$$

This set includes the job ads from the standard recommendation set $J_s(i)$ that are within the top K based on their ranking R_{sij} . The personalized recommendation set until the top K for an individual i is defined as:

$$R_p^{(K)}(i) = \{j \in J_p(i) \mid R_{pij} \leq K\}$$

This set includes job ads from the personalized recommendation set $J_p(i)$ that are within the top K based on their ranking R_{pij} . The overlap set until the top K is defined as the set of job ads that are common between the top K standard and personalized recommendation sets for individual i :

$$R_{\text{overlap}}^{(K)}(i) = R_s^{(K)}(i) \cap R_p^{(K)}(i)$$

We now introduce $\text{Overlap@K}(i)$, which is defined as the proportion of overlapping vacancies between the personalized *top-K* and the standard *top-K* for each job seeker:

$$\text{Overlap@K}(i) = \frac{|R_{\text{overlap}}^{(K)}(i)|}{\min(|\mathcal{R}_p^K(i)|, |\mathcal{R}_s^K(i)|)} \quad (2.17)$$

We observe that the Overlap@15 between the two rankings is substantial: the median overlap is 80% (12 ads in common), and only 25% of job seekers have fewer than 7 ads in common between the two top-15s. This excessive overlap was partially caused by the implementation issue, which led to recommendations being sorted by date instead of relevance for both algorithms. Consequently, the most recent vacancies were at the top of both rankings, increasing the likelihood of finding identical ads in both rankings. To further illustrate this, we calculated the potential rankings, i.e., the rankings that job seekers would have encountered if the vacancies had been ranked by relevance rather than

by date. By doing so, we can compare the realized overlaps with the potential overlaps (which would have been observed without the implementation issue). We discover that realized overlaps are larger than potential overlaps: the proportion of ads in common between the two rankings would have been lower if the rankings had been sorted by relevance. For instance, we find that realized Overlap@10 is greater than potential Overlap@10 for 75% of job seekers. The distributions of realized and potential overlaps are represented on figure H.13.

6 Results of the large scale experiment

In this section, we present the findings from our large-scale experiment designed to evaluate the effectiveness of personalized job recommendations. All reported effects are Intent-To-Treat (ITT) estimates. Specifically, as detailed in Table 3.1, a portion of job seekers assigned to the full and partial treatment groups did not comply to their assigned treatment—either by disconnecting from the platform before completing the weight entry task or by inadvertently being reassigned to the control group. Consequently, our analysis estimates the effect of being assigned to the treatments rather than the effect of actually receiving the treatments. To avoid overcomplicating the presentation of our results, we provide additional analyses of the Average Treatment Effect (ATE) on compliers in the appendix, demonstrating that these findings align with the main results.

As detailed in the data section, our analysis focuses on three key outcomes to assess user engagement and adoption of job recommendations at the first connection to the online platform: (1) the number of clicks on recommended job ads, which serves as a measure of initial interest; (2) the number of intentions to apply, indicated by clicks on the “apply” button, reflecting a deeper level of engagement; and (3) the number of actual job applications submitted, which is the most policy-relevant outcome as it directly correlates with the likelihood of successful job placements. Additionally, we consider (4) the total number of connections to the platform over time, which reflects the broader interest in the platform itself.

6.1 Effect of personalized recommendations compared to standard recommendations

In this section, we assess the impact of receiving job recommendations generated by the personalized algorithm, which uses individualized weights, in contrast to recommendations produced by the standard algorithm. This comparison focuses on the outcomes observed in the full treatment group versus the partial treatment group.

For each outcome, we estimate via OLS the following model:

$$Y_i = \alpha + \beta \text{FullTreatment}_i + \mu Z_i + \varepsilon_i \quad (2.18)$$

where Y_i represents the outcome of interest for job seeker i , FullTreatment_i is a binary variable that equals 1 if the job seeker was assigned to the full treatment group and 0 if the job seeker was assigned to the partial treatment group, and Z_i is a vector of controls accounting for the stratified randomization design. To gain a deeper understanding of the impact of personalized recommendations, we distinguish between clicks on all recommendations and clicks on the subset of overlapping recommendations. Overlapping recommendations are those job ads that are recommended by both the standard and personalized algorithms. As explained in the previous section, differentiating between these two types of clicks allows us to isolate the effect of newly introduced job ads by the personalized algorithm. By comparing the outcomes for all recommendations (\mathcal{R}_{seen}) to those for only overlapping recommendations ($\mathcal{R}_{overlap}$), we can determine whether the observed treatment effects are driven by the introduction of new job ads that were not part of the standard recommendations, or if they simply reflect a reallocation of interest from the standard ads to those prioritized by the personalized algorithm.

Table 2.8: Impact of the personalization treatment on recommendations appreciation and adoption

Dependent variable	Clicks on recommendations		Intentions to apply on recommendations		Applications on recommendations		Connections to online platform
	# clicks on:		# intents to apply on:		# applications on:		
	\mathcal{R}_{seen} (1.A)	$\mathcal{R}_{overlap}$ (1.B)	\mathcal{R}_{seen} (2.A)	$\mathcal{R}_{overlap}$ (2.B)	\mathcal{R}_{seen} (3.A)	$\mathcal{R}_{overlap}$ (3.B)	
Treatment status (ref. partial treatment):							
Full treatment	0.067 (0.022)**	-0.008 (0.019)	0.015 (0.007)**	0.003 (0.006)	0.006 (0.003)*	0.001 (0.003)	0.019 (0.011*)
Strata fixed effects	yes	yes	yes	yes	yes	yes	yes
N.Obs.	20 845	20 845	20 845	20 845	20 845	20 845	20 845
Partial treatment mean	0.484	0.423	0.072	0.065	0.026	0.024	0.961

Note: This table shows ITT results from equation 2.18. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

Table 2.8 shows that job seekers in the full treatment group clicked on recommendations slightly more frequently than those in the partial treatment group. The full treatment resulted in an increase of 0.067 clicks on average, representing a 13.8% increase over the partial treatment group's mean of 0.484 clicks. In terms of intentions to apply, the full treatment group exhibited a significant increase of 0.015 intentions on average, a 20.8% increase from the partial treatment group's mean of 0.072. Similarly, the number of ac-

tual applications submitted also showed a slight increase of 0.006, reflecting a 23.1% rise from the partial treatment group's mean of 0.026 applications. Additionally, we observe that the full treatment group experienced a slight increase in the number of connections to the online platform (0.019), representing a modest 2.0% rise from the partial treatment group's mean of 0.961 connections. We also provide in the appendix the estimation of the full treatment effect on the sample of the full respondents (those who completed the weight entry task), which are consistent with these results, see Table I.8.

Interestingly, when we restrict our analysis to overlapping recommendations, there is no significant difference between the outcomes of job seekers exposed to personalization and those who were not. This indicates that the positive average treatment effects identified are primarily driven by the extensive margin —namely, the new job ads introduced by the personalized algorithm. The personalized recommendations did not simply cause job seekers to shift their attention from standard ads to those emphasized by the personalized algorithm, but rather led to additional engagement with new, relevant job opportunities.

While these increases are statistically significant, they are relatively small in magnitude. The impact on application for example is only 0.006, the impact on the intention to apply on recommendations is 0.015 and the impact on clicks is 0.067. However, in relative terms, these impacts are more substantial: the impact on clicks represents a 14% increase compared to the control group, the impact on intentions to apply and actual applications represents respectively a 21% and 23% increase compared to the control group. In addition, it is important to consider the potential influence of the misconfiguration in the job ad sorting mechanism, which sorted recommendations by their posting date rather than relevance. This issue likely caused job seekers to interact more with recently posted ads, irrespective of their relevance to individual preferences. As a result, the positive effects of the weight-based personalization treatment may be understated. Job seekers may have been drawn to newer ads out of curiosity, thereby diluting the measurable impact of the personalized recommendations and leading to an underestimation of the true benefits of this treatment.

6.2 Evaluating the impact of the stated preference data collection itself

We now investigate the impact of completing the weight entry task before receiving job recommendations, compared to receiving recommendations directly without entering any weights. This comparison is made between the partial treatment group and the control group.

Our results are estimated via OLS using the following specification:

$$Y_i = \alpha + \beta \text{PartialTreatment}_i + \mu Z_i + \varepsilon_i, \quad (2.19)$$

where $\text{PartialTreatment}_i$ is a binary variable indicating whether individual i was assigned to the partial treatment group ($\text{PartialTreatment}_i = 1$) or the control group ($\text{PartialTreatment}_i = 0$). The outcome variable Y_i represents one of the three outcomes of interest for individual i , and Z_i is a vector controlling for the stratified randomization design. The coefficient β captures the ITT estimate, reflecting the impact of being assigned to the partial treatment group relative to the control group. For individuals in the partial treatment group who exited the experiment early—before seeing any recommendations because they chose not to complete the weight entry task—we assign a value of 0 to all outcome variables.

Table 2.9 presents the results of being assigned to the partial treatment group across the three key outcome variables. The estimated treatment effect on the number of clicks is -0.267 and significant at the 1% level. Given the control group’s mean of 0.753 clicks, this represents a substantial reduction in engagement. The effect on intentions to apply is -0.048 , also statistically significant at the 1% level. Considering the control group mean of 0.121, this reduction is relatively large. The treatment effect on actual applications submitted is -0.020 , statistically significant at the 5% level and representing a 43% decrease. Moreover, the partial treatment led to a decrease of -0.334 connections to the online platform, a 25.7% drop relative to the control group’s mean of 1.297 connections. Table I.7 in the appendix provides the ATE estimated on compliers, confirming the same negative effects of the partial treatment relative to the control group.

Table 2.9: Impact of being assigned to the partial treatment group

Dependent variable	Clicks on recommendations (1)	Intentions to apply on recommendations (2)	Applications on recommendations (3)	Connections to online platform (4)
Treatment status (ref. control):				
Partial treatment	-0.267 (0.023)***	-0.048 (0.008)***	-0.020 (0.005)**	-0.334 (0.020)***
Strata fixed effects	yes	yes	yes	yes
N. Obs.	20 995	20 995	20 995	20 995
Control mean	0.753	0.121	0.046	1.297

Note: This table shows ITT results from equation 2.19. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

These findings indicate that assignment to the partial treatment group leads to a reduction in recommendation interest and adoption. A particularly notable finding is the significant reduction in the number of connections to the platform itself. Since the plat-

form offers services that are not as easily accessible on the general PES website, this decline is particularly concerning. Fewer platform visits imply fewer opportunities for job seekers to find suitable job matches, which could diminish the overall effectiveness of the intervention and harm labor market outcomes.

One key factor behind this decline could be that the weight entry task set higher expectations for the relevance of the recommendations. Job seekers, having invested effort in expressing their job preferences, might have anticipated more personalized and relevant job matches. When these expectations were not met, they may have become disillusioned with the platform.

However, it is important to emphasize that job seekers in the partial treatment group probably did not feel misled. Although their preferences were not used in generating the recommendations, participants may have assumed that their preferences were considered, given the weight allocation task they completed. The dissatisfaction they expressed was more likely due to the overall poor relevance of the recommendations they received rather than a belief that their stated preferences were ignored. In this way, their disappointment could stem from the gap between the heightened expectations created by the task and the quality of the job recommendations.

Additionally, the misconfiguration in the sorting mechanism that ordered job ads by posting date rather than relevance may have further exacerbated the disappointment effect for partially treated job seekers. Those in the partial treatment group, who completed the weight entry task, might have been more disappointed when their expectations for highly personalized and relevant recommendations were not met due to this sorting error. Consequently, the negative effects observed in this group might reflect not just the heightened expectations associated with the weight entry task but also the frustration of encountering recommendations that were not the most relevant available.

6.3 Overall effects of personalized job recommendations using stated preferences

In this section, we examine whether the introduction of personalized job recommendations, tailored according to job seekers' stated preferences, enhances recommendation appreciation and adoption compared to the standard recommender system. Specifically, we compare the outcomes of job seekers assigned to the full treatment group with those in the control group. These results, when combined with those from the previous two sections, illustrate the overall net effect of the intervention as jointly conducted.

To assess the impact of the full treatment, we estimate via OLS the following model:

$$Y_i = \alpha + \beta \text{FullTreatment}_i + \mu Z_i + \epsilon_i, \quad (2.20)$$

where FullTreatment_i is a binary variable that equals 1 if individual i was assigned to the full treatment group and 0 if assigned to the control group. The dependent variable Y_i represents one of the three outcomes of interest (clicks, intentions to apply, or applications) for individual i . The vector Z_i includes control variables reflecting the stratified randomization design. The coefficient β captures the average treatment effect of being assigned to the full treatment group relative to the control group. Note that for individuals in the full treatment group who exited the experiment early, before seeing any recommendations because they chose not to complete the weight entry task, we assign a value of 0 to all outcome variables.

As expected from the results in the two previous sections, Table 2.10 indicates that the assignment to the full treatment group significantly reduces the number of interactions across all outcomes compared to the control group. Specifically, the full treatment leads to a decrease of 0.2 clicks on job recommendations, a 26% reduction relative to the control group's average of 0.753 clicks. Similarly, the number of intentions to apply decreases by 0.033, representing a 27% reduction, and the number of actual job applications submitted declines by 0.014, or a 30% reduction compared to the control group's baseline of 0.046 applications. In addition to these outcomes, the full treatment also significantly decreases the number of connections to the online platform by 0.315, a reduction of nearly 24% from the control group's mean of 1.297 connections. Table I.6 in the appendix provides the ATE estimated on compliers, confirming the same negative effects of the full treatment relative to the control group.

Table 2.10: Impact of being assigned to the full treatment group

Dependent variable	Clicks on recommendations (1)	Intentions to apply on recommendations (2)	Applications on recommendations (3)	Connections to online platform (4)
Treatment status (ref. control):				
Full treatment	-0.200 (0.024)***	-0.033 (0.008)***	-0.014 (0.005)**	-0.315 (0.020)***
Strata fixed effects	yes	yes	yes	yes
N.Obs.	20 832	20 832	20 832	20 832
Control mean	0.753	0.121	0.046	1.297

Note: This table shows results from equation 2.20. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

Conclusion

This study set out to explore the effectiveness of personalized job recommendations based on job seekers' explicitly stated preferences. The stated preference data were collected using a constant-sum method, where job seekers allocated points among various job attributes: wage, commuting distance, working hours, occupation, and contract type. This method ensured that job seekers expressed their priorities clearly and that these preferences could be directly used to personalize the matching algorithm.

The reliability of the stated preferences was validated in a pilot study, where job seekers demonstrated consistency in their responses. Analysis of these stated preferences revealed significant deviations from the standard weights used in the PES matching algorithm. Job seekers placed different levels of importance on attributes such as wage and commuting distance compared to the generic, one-size-fits-all weights typically applied by the PES. This finding underscores the limitations of a uniform approach to job matching and highlights the potential benefits of a more individualized system.

By conducting a large-scale randomized experiment, we evaluated whether a more personalized matching algorithm could enhance job seekers' engagement with job ads, as measured by their clicks, intentions to apply, and actual applications. Our findings indicate the personalized matching algorithm led to an increase in engagement. This increase was primarily driven by new job ads introduced by the personalized algorithm. Interestingly, these effects were observed without a simple shift in attention from standard recommendations, indicating that personalization contributed to expanding the range of considered opportunities. However, the overall effect of personalization was negative. The requirement to personalize weights appears to have heightened job seekers' expectations for highly relevant recommendations. The challenges of personalization were further compounded by an unforeseen technical misconfiguration, which sorted job ads by date rather than relevance on our online platform for all treatment arms⁷ (including the control), leaving the internal validity of the experiment intact. This issue likely amplified disappointment among treated job seekers, resulting in increased negative treatment effects. Therefore, we believe we estimated the lower bound of the treatment effect, as it is unlikely that the impact of personalization could be worse than what we observed without this issue.

This research contributes to the literature on job recommender systems by highlighting the delicate balance between personalization and user experience. The negative impacts observed in both the partial and full treatment groups underscore the importance

⁷Recognizing these complexities, we are planning to launch a new experiment in October 2024 to address the issues identified in this study. The upcoming experiment will correct the technical misconfiguration and reassess the impact of personalized recommendations.

of ensuring that any additional cognitive burdens imposed by personalization tasks are justified by a corresponding increase in the relevance and quality of recommendations.

This study, however, has several limitations. First, the job matching algorithm relied on the registered search criteria of the job seekers, which might not always be up-to-date. This is a recurrent issue at the PES, as job seekers can change their preferences or adapt their search parameters throughout their unemployment spell [Marinescu and Skandalis, 2021]. An alternative approach could involve allowing job seekers to adjust the absolute values of their search parameters, not just their relative importance. Additionally, it is worth noting the solution proposed by Bächli et al. [2024], who constructed a recommendation algorithm that enables job seekers to weight the importance of factors such as their previous occupation and profile, potentially capturing changes in their job search behavior more effectively. Second, the personalization process relied solely on stated preferences, which might not fully capture the nuances of job seekers' true priorities in practice. Stated preferences can sometimes diverge from actual behavior due to various factors. In this regard, an essential next step is to verify whether the stated preferences collected through point allocation are consistent with the organic job search behavior of job seekers. This comparison between stated and revealed preferences could provide valuable insights into the effectiveness of the preference elicitation process and help refine algorithms to better align with actual job seeker behavior. Third, the technical misconfiguration that affected the sorting of job ads likely confounded our ability to measure the true impact of personalization accurately. The reduced engagement we observed might have been partially driven by this misconfiguration rather than the personalization itself. Finally, our study did not consider the long-term effects of personalization on job seekers' outcomes. It is unclear whether the initial variations in engagement would translate into varying employment rates over time. Indeed, the relatively low participation in the experiment resulted in reduced statistical power to detect effects on employment outcomes.

One area for future exploration could involve designing an adaptive system that tailors the user interaction effort to individual characteristics, recognizing that users differ in their predisposed levels of effort. By dynamically adjusting the complexity of the personalization process, such a system could motivate users to invest an optimal amount of effort in expressing their preferences. Another promising direction for future research is the development of an interactive platform where job seekers can actively adjust the weights they assign to job attributes and immediately see how these changes affect the job recommendations they receive. Such a system would allow job seekers to better understand how their preferences shape the matching process, potentially leading to more engagement with the platform. While this approach could enhance the overall user ex-

perience and personalization process, it was not feasible in our study due to technical constraints. Further research could also examine whether implicit data, such as users' past behaviors and interactions, could be useful to personalize job attribute importance within the algorithm. This approach has the potential to reduce the burden on job seekers by eliminating the need for explicit preference elicitation. However, using implicit data is not straightforward. People's behavior online can be inconsistent—clicks may result from curiosity, habit, or chance, rather than genuine interest. As a result, it can be difficult for an algorithm to accurately interpret these actions as true preferences. Additionally, unlike stated preferences, implicit data is not constrained to a specific set of choices, making it challenging to estimate a full range of preferences for each individual.

References

- Steffen Altmann, Anita M Glenny, Robert Mahlstedt, and Alexander Sebald. The direct and indirect effects of online job search advice. 2022.
- Mirjam Bächli, H el ene Benghalem, Doriana Tinello, Damaris Aschwanden, Sascha Zuber, Matthias Kliegel, Michele Pellizzari, and Rafael Lalive. Ranking occupations by their proximity to workers' profiles. Swiss Journal of Economics and Statistics, 160(1):8, 2024.
- Abhijit V Banerjee and Gaurav Chiplunkar. How important are matching frictions in the labor market? experimental & non-experimental evidence from a large indian firm. Journal of Development Economics, page 103330, 2024.
- Luc Behaghel, Sofia Dromundo, Marc Gurgand, Yagan Hazard, and Thomas Zuber. The potential of recommender systems for directing job search: A large-scale experiment. 2024.
- Svetlin Bostandjiev, John O'Donovan, and Tobias H ollerer. Tasteweights: a visual interactive hybrid recommender system. In Proceedings of the sixth ACM conference on Recommender systems, pages 35–42, 2012.
- Paul A Bottomley and John R Doyle. Comparing the validity of numerical judgements elicited by direct rating and point allocation: Insights from objectively verifiable perceptual tasks. European Journal of Operational Research, 228(1):148–157, 2013.
- Paul A Bottomley, John R Doyle, and Rodney H Green. Testing the reliability of weight elicitation methods: direct rating versus point allocation. Journal of Marketing Research, 37(4):508–513, 2000.
- Mich ele B elot, Philipp Kircher, and Paul Muller. How wage announcements affect job search: A field experiment. IZA Discussion Papers, 11814, September 2018.
- Mich ele B elot, Philipp Kircher, and Paul Muller. Do the long-term unemployed benefit from automated occupational advice during online job search? SSRN Electronic Journal, 01 2022. doi: 10.2139/ssrn.4178928.
- Peter A Diamond and Jerry A Hausman. Contingent valuation: is some number better than no number? Journal of economic perspectives, 8(4):45–64, 1994.
- John R Doyle, Rodney H Green, and Paul A Bottomley. Judging relative importance: direct rating and point allocation are not equivalent. Organizational behavior and human decision processes, 70(1):65–72, 1997.

- Sara Drenner, Shilad Sen, and Loren Terveen. Crafting the initial user experience to achieve community goals. In Proceedings of the 2008 ACM conference on Recommender systems, pages 187–194, 2008.
- Brian Feld, AbdelRahman Nagy, and Adam Osman. What do jobseekers want? comparing methods to estimate reservation wages and the value of job attributes. Journal of Development Economics, 159:102978, 2022. ISSN 0304-3878. doi: <https://doi.org/10.1016/j.jdeveco.2022.102978>. URL <https://www.sciencedirect.com/science/article/pii/S0304387822001201>.
- Mauricio N Freire and Leandro N de Castro. e-recruitment recommender systems: a systematic review. Knowledge and Information Systems, 63:1–20, 2021.
- Gerald Häubl and Valerie Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. Marketing science, 19(1):4–21, 2000.
- Roland Helm, Armin Scholl, Laura Manthey, and Michael Steiner. Measuring customer preferences in new product development: comparing compositional and decompositional methods. International Journal of Product Development, 1(1):12–29, 2004.
- Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS), 20(4):422–446, 2002.
- Jon A Krosnick and Duane F Alwin. A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. Public Opinion Quarterly, 52(4): 526–538, 1988.
- Thomas Le Barbanchon, Roland Rathelot, and Alexandra Roulet. Gender differences in job search: Trading off commute against wage. The Quarterly Journal of Economics, 136(1):381–426, 2021.
- Thomas Le Barbanchon, Lena Hensvik, and Roland Rathelot. How can ai improve search and matching? evidence from 59 million personalized job recommendations. Technical report, Working Paper, 2023.
- Danielle Li, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Technical report, National Bureau of Economic Research, 2020.
- Nicole Maestas, Kathleen J Mullen, David Powell, Till Von Wachter, and Jeffrey B Wenger. The value of working conditions in the united states and implications for the structure of wages. American Economic Review, 113(7):2007–2047, 2023.

- Charles F Manski, Kenneth I Wolpin, and Elke U Weber. Analysis of choice expectations in incomplete scenarios. Elicitation of Preferences, pages 49–72, 2000.
- Ioana Marinescu and Daphné Skandalis. Unemployment insurance and job search behavior. The Quarterly Journal of Economics, 136(2):887–931, 2021.
- Kevin Marsh, Maarten IJzerman, Praveen Thokala, Rob Baltussen, Meindert Boysen, Zoltán Kaló, Thomas Lönngren, Filip Mussen, Stuart Peacock, John Watkins, and Nancy Devlin. Multiple criteria decision analysis for health care decision making—emerging good practices: Report 2 of the ispor mcda emerging good practices task force. Value in Health, 19(2):125–137, 2016. ISSN 1098-3015. doi: <https://doi.org/10.1016/j.jval.2015.12.016>. URL <https://www.sciencedirect.com/science/article/pii/S1098301515300152>.
- Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. American Economic Review, 107(12):3722–59, 2017a.
- Alexandre Mas and Amanda Pallais. Valuing alternative work arrangements. American Economic Review, 107(12):3722–3759, 2017b.
- John A McCarty and Larry J Shrum. The measurement of personal values in survey research: A test of alternative rating procedures. Public Opinion Quarterly, 64(3):271–298, 2000.
- Denis Parra and Peter Brusilovsky. User-controllable personalization: A case study with selffusion. International Journal of Human-Computer Studies, 78:43–67, 2015.
- Pearl Pu, Li Chen, and Rong Hu. Evaluating recommender systems from the user’s perspective: survey of the state of the art. User Modeling and User-Adapted Interaction, 22:317–355, 2012.
- Mandy Ryan, Diane Alison Scott, C Reeves, A Bate, Edwin Roland Van Teijlingen, Elizabeth M Russell, M Napper, and CM Robb. Eliciting public preferences for healthcare: a systematic review of techniques. Health technology assessment (Winchester, England), 5(5):1–186, 2001.
- Kirsten Swearingen and Rashmi Sinha. Beyond algorithms: An hci perspective on recommender systems. In ACM SIGIR 2001 workshop on recommender systems, volume 13, pages 1–11, 2001.
- Koert Van Ittersum, Joost ME Pennings, Brian Wansink, and Hans CM Van Trijp. The validity of attribute-importance measurement: A review. Journal of Business Research, 60(11):1177–1190, 2007.

Marieke GM Weernink, Sarah IM Janus, Janine A Van Til, Dennis W Raisch, Jeannette G Van Manen, and Maarten J IJzerman. A systematic review to identify the use of preference elicitation methods in healthcare decision making. Pharmaceutical medicine, 28: 175–185, 2014.

Matthew Wiswall and Basit Zafar. Preference for the workplace, investment in human capital, and gender. The Quarterly Journal of Economics, 133(1):457–507, 2017.

World Bank. The use of advanced technology in job matching platforms: Recent examples from public agencies. 2023. URL <https://thedocs.worldbank.org/en/doc/ceb5c5792ad0d874e9b1c3cc71362f46-0460012023/original/Digital-Job-Matching-Platforms-S4YE-Draft-Note-for-Discussion.pdf>.

Noorul Hassan Zardari, Kamal Ahmed, Sharif Moniruzzaman Shirazi, and Zulkifli Bin Yusop. Weighting methods and their effects on multi-criteria decision making model outcomes in water resources management. Springer, 2014.

Appendices

A The PES matching algorithm

This section outlines the basic functioning of the PES matching algorithm, though its actual implementation is far more complex in practice.

A.1 General formula

The overall matching score is computed as follows:

$$s(i, j) = f_{ij} \sum_{k=1}^K \frac{w_k}{\sum_{k=1}^K w_k} a_k(i, j)$$

with $a(i, j)$ the matching score between vacancy j and jobseeker i , w_k the weight assigned to criterion k and $a_k(i, j)$ the adequacy measure between the vacancy attribute k and the jobseeker characteristic k . $f_{ij} \in \{0, 1\}$ is a filter variable which is equal to 0 if at least one of the following adequacy measure is null: maximum commuting distance, languages and driving license.

A.2 Adequacy measures for each attribute

*Skills

$$a_{\text{skills}}(i, j) = \frac{1}{|S_j|} \sum_{s \in S_j} \max_{s' \in S_i} A(s, s')$$

with:

- S_i : jobseeker's skills
- S_j : skills required by the recruiter
- $A(., .)$: affinity matrix between skills

*Occupation

$$a_{\text{occupation}}(i, j) = A(O_i, O_j)$$

with:

- O_i : occupation targeted by the jobseeker
- O_j : occupation proposed by the recruiter

- $A(.,.)$: affinity matrix between occupations

*Wage

$$a_{\text{wage}}(i, j) = \begin{cases} 1 & \text{if } s_i \leq s_j \\ 1 - \frac{s_i - s_j}{s_i - B} & \text{if } B \leq s_j < s_i \\ 0 & \text{otherwise} \end{cases}$$

with:

- s_i : reservation wage
- s_j : wage proposed by recruiter
- $B := 0.89 \times s_i$

*Working experience

$$a_{\text{experience}}(i, j) = \begin{cases} 1 & \text{if } e_j \leq e_i \\ 1 - \frac{e_j - e_i}{B - e_i} & \text{if } e_i \leq e_j \leq B \\ 0 & \text{otherwise} \end{cases}$$

with:

- e_i : jobseeker's experience, in months
- e_j : months of experience required by the recruiter
-

$$B := \begin{cases} e_i + 25, & \text{if } e_i \leq 36 \\ (e_i \times 4) + 1 & \text{otherwise} \end{cases}$$

*Working hours

$$a_{\text{working hours}}(i, j) = \begin{cases} 1 & \text{if } t_{\min} \leq t_o \leq t_{\max} \\ E_{\min} & \text{if } (t_o \leq t_{\max}) \& (t_o \leq t_{\min}) \\ E_{\max} & \text{if } (t_o \geq t_{\min}) \& (t_o \geq t_{\max}) \\ \min(E_{\min}, E_{\max}) & \text{otherwise} \end{cases}$$

with:

- t_i, \min : minimum weekly duration accepted by the jobseeker

- $t_{i, max}$: maximum weekly duration accepted by the jobseeker
- t_j : weekly duration proposed by the recruiter
- $E_{max} = 1 \left\{ (t_0 \geq t_{max}) \& (t_0 - t_{max} \leq 2) \right\} \left(1 - \frac{t_0 - t_{max}}{2} \right)$
- $E_{min} = 1 \left\{ (t_0 \leq t_{min}) \& (t_0 - t_{min} \geq -2) \right\} \left(1 - \frac{t_{min} - t_0}{2} \right)$

*Commuting distance

$$a_{\text{distance}}(i, j) = \begin{cases} 1 & \text{if } d_{ij} \leq d_{max} \\ 1 - \frac{d_{ij} - d_{max}}{B - d_{max}} & \text{if } d_{max} \leq d_{ij} \leq B \\ 0 & \text{if } d_{ij} > B \end{cases}$$

with:

- $d_{ij} := \text{distance}(\text{jobseeker's location, company's location})$
- d_{max} : maximum commuting distance accepted by the jobseeker
- $B := \max(1.3 \times d_{max}, 10)$

*Duration and type of contract

$$a_{\text{type of contract}}(i, j) = A(C_i, C_j)$$

with:

- C_i : type of contract targeted by the jobseeker
- C_j : type of contract proposed by the recruiter
- $A(., .)$: affinity matrix between contract types

*Education

$$a_{\text{diploma}}(i, j) = A(D_i, D_j)$$

with:

- D_i : diploma(s) held by the jobseeker
- D_j : diploma(s) required by the recruiter
- $A(., .)$: affinity matrix between diploma types

*Languages

$$a_{\text{languages}}(i, j) = \frac{|L_i \cap L_j|}{|L_j|}$$

with:

- L_i : jobseeker's spoken language(s)
- L_j : language(s) preferred (not mandatory) or required (mandatory) by recruiter

*Driving license

$$a_{\text{driving}}(i, j) = \frac{1}{|D_j|} \sum_{d \in D_j} \max_{d' \in D_i} A(d, d')$$

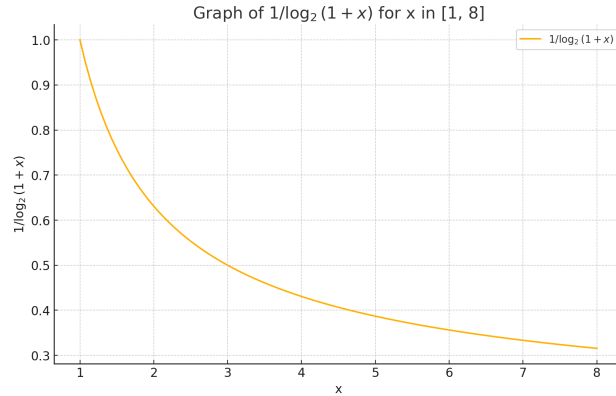
with:

- D_i : jobseeker's driving license(s)
- D_j : driving license(s) required by recruiter
- $A(., .)$: affinity matrix between driving license types

B Metrics to compare lists of recommended ads

B.1 Weights used in the DCG

Figure B.1: Inverse of $\log_2(1+x)$



B.2 Alternative metrics

The Ranking Biased Overlap (RBO) The RBO, introduced by [Webber et al. \[2010\]](#), measures the cardinal of the overlap of two sublists at a given depth. The RBO between lists S and T is computed as follows:

$$RBO(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} \cdot A_d \quad (\text{B.1})$$

where d is the depth in the list, p is a weighting factor (between 0 and 1) and A is the number of intersecting items at depth d divided by the depth d itself. Thanks to p , differences at the top of the list ought to be given more weight than differences further down. A RBO equal to 1 means that the two lists are exactly the same, whereas a RBO equal to 0 means that they are disjoint. The advantage of RBO is that it can be used to compare uneven lists.

Kendall tau distance and Kendall tau correlation The Kendall tau distance is a metric that computes the number of concordant pairs between two sorted lists. A pair is concordant if the two items appear in the same order in each list. If element x is placed before element y in both list 1 and in list 2, they will count as a concordant pair, whatever the distance that separates them between the two lists. After having computed the number of concordant pairs, the Kendall tau sum is normalized by $\binom{n}{2}$, i.e. the total number of pairs that could be made. Another version of the Kendall tau measure is the Kendall tau coefficient of correlation. There, we count the number of concordant pairs (denoted N_c) and

the number of discordant ones (denoted N_d) and obtain the τ coefficient of correlation :

$$\tau = \frac{N_c - N_d}{\binom{2}{n}} \quad (\text{B.2})$$

The Kendall tau coefficient ranks between -1 and 1. The more it is close to 1, the more the lists are similar. When the coefficient is equal to -1, that means that the two lists are inverted in relation to each other. The disadvantage of Kendall tau distance is that it can be only computed on lists of the same length. Then, the Kendall Tau distance and coefficient do not factor in the distance of dissimilarity between two elements. For instance, if considering two elements x and y that are located in the same order in the two lists (x before y) but not at the same distance (x just before y in list 1 and x for example 8 ranks before y in list 2), they will still count for the same value in the Kendall tau score. Finally, the Kendall tau metric gives exactly the same weight to concordant or discordant pairs, whatever the depth in the list.

C User flows on the online interface

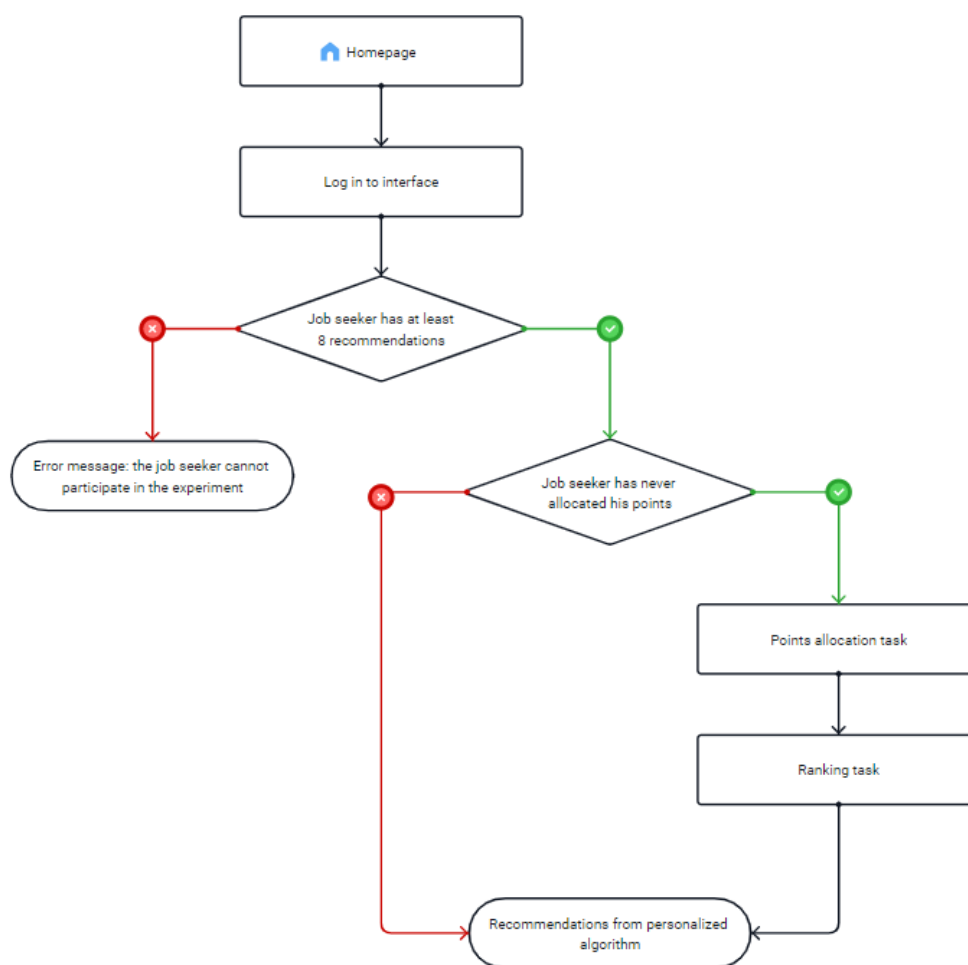


Figure C.2: User flow - pilot study

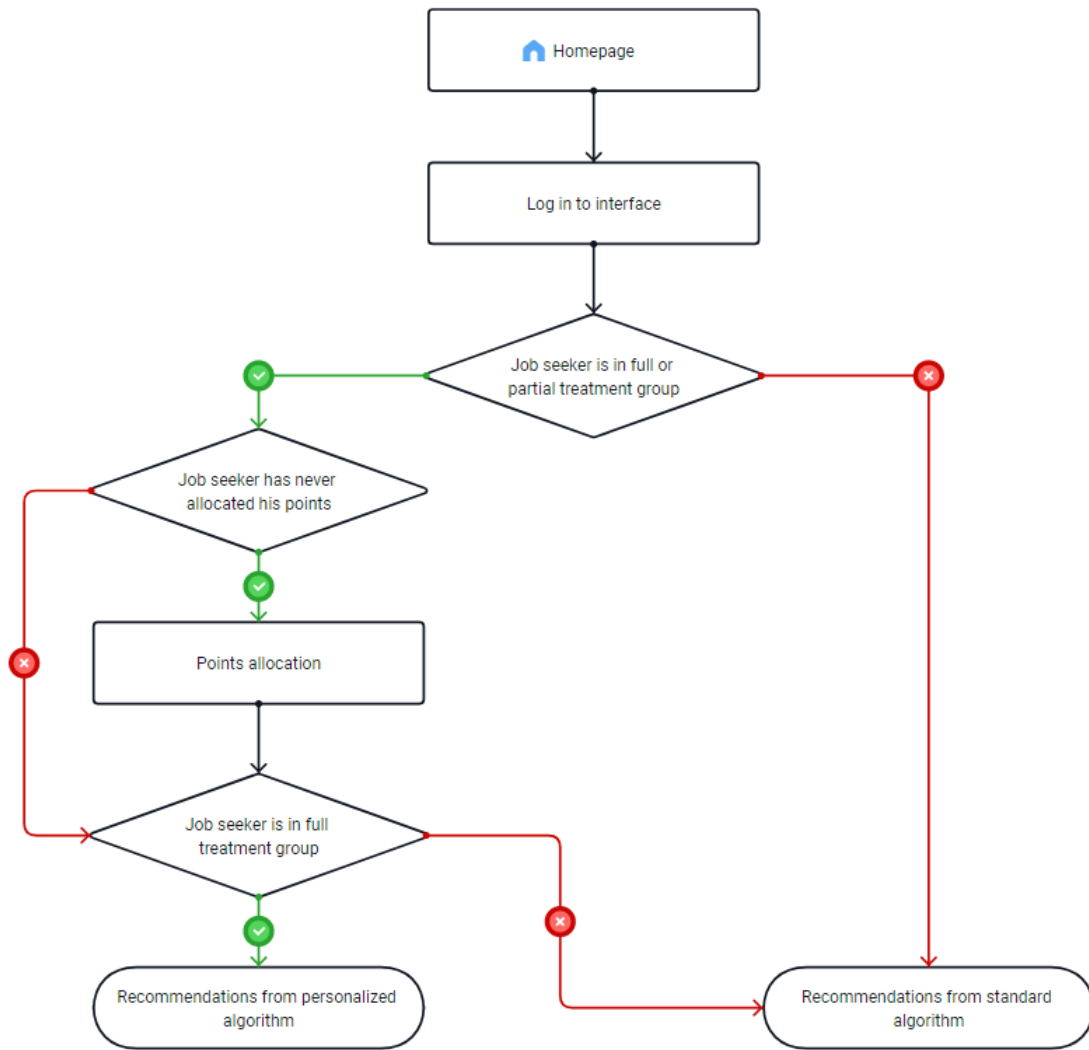


Figure C.3: User flow - large scale experiment

D Screenshots of the user interface, translated from French

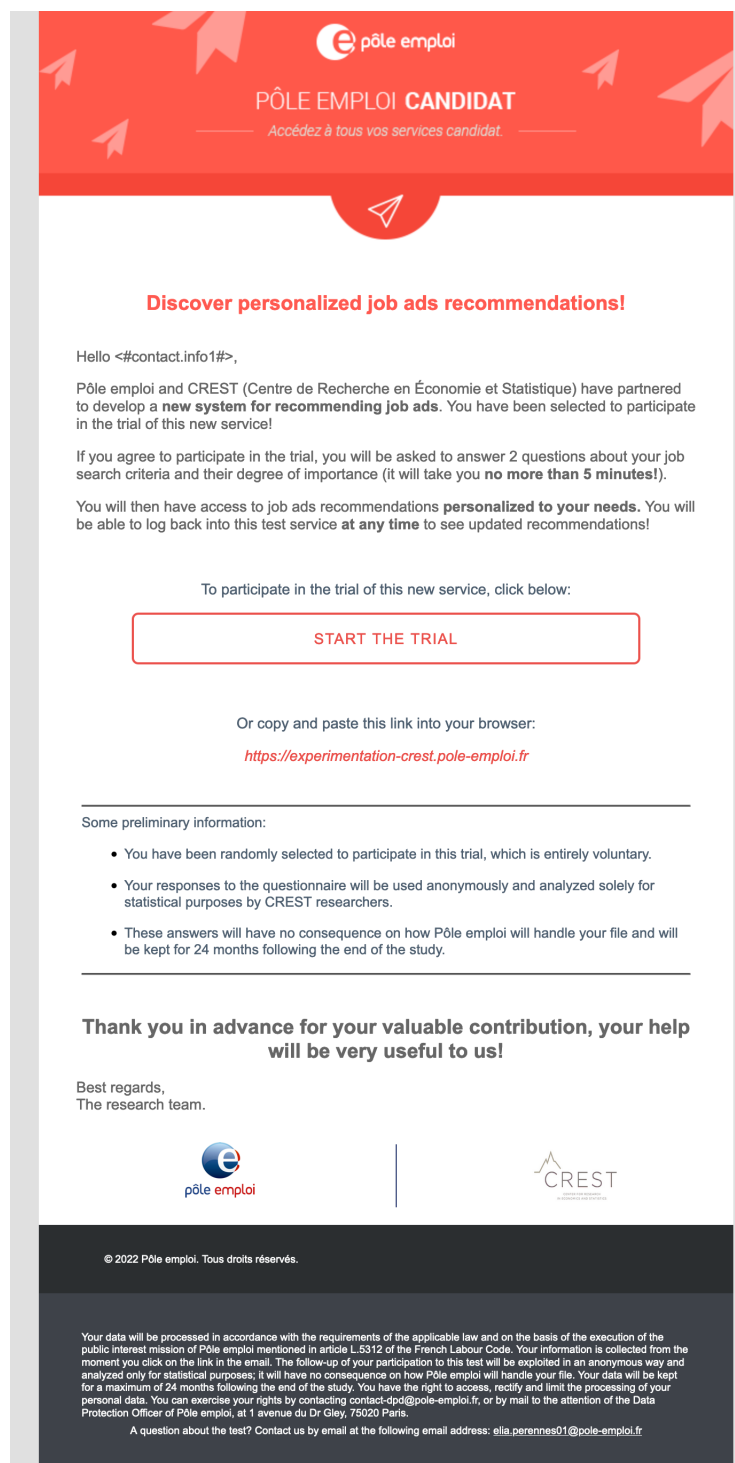


Figure D.4: Pilot study, invitation email



Discover personalized job ad recommendations!

Hello <#contact.info1#>,

Pôle emploi and the CREST (Center for Research in Economics and Statistics) have joined forces to develop a new job ad suggestion service. You have been selected to test this new service!

By participating in this test, you will have access for a period of 2 months to personalized job offer recommendations, which will be updated each time you visit. The purpose of this service is to facilitate your job search by providing you with recommendations that are particularly well-suited to your search with just one click.

As part of this test, we may also ask you questions about your priorities concerning your job search.

To participate in the trial of this new service, click below:

START THE TEST →

Or copy and paste this link into your browser :

<https://experimentation-crest.pole-emploi.fr/<#contact.info2#>>

Some preliminary information:

- Your navigation data on the service will be used anonymously and analyzed solely for statistical purposes by the researchers at CREST.
- This data will have no impact on the management of your file by Pôle emploi, and no data that allows your identification will be retained.

Thank you in advance for your valuable collaboration, your help will be very useful to us!

Sincerely,
The research team.



© 2022 Pôle emploi. All rights reserved.

Your data will be processed in compliance with the requirements of applicable law and based on the public interest mission of Pôle emploi mentioned in Article L.5312 of the Labor Code. Your information is collected from the moment you click on the link in the email. The monitoring of your participation in this test will be used anonymously and analyzed solely for statistical purposes; this will have no impact on the management of your file by Pôle emploi. No data that allows your identification will be retained. You have the right to access, rectify, and limit the processing of your personal data. You can exercise your rights by contacting contact-dpd@pole-emploi.fr, or by mail addressed to the Data Protection Officer of Pôle emploi, at 1 avenue du Dr Gley, 75020 Paris.

Have a question about the test? Contact us by email at the following address: elia.perennes01@pole-emploi.fr

Figure D.5: Large experiment, invitation email

Help us create the job recommendation system of the future!


A quick, personalized, anonymous and safe test:

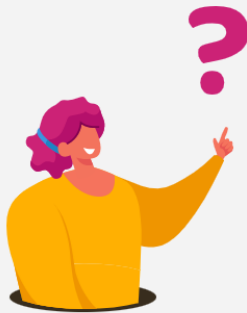
This test requires about 5 minutes to complete. The objective is to propose personalized job ads that match each person's priorities.

To access the test, log in with your Pôle emploi credentials by clicking on the "Log in with Pôle emploi" button. This secured device is essential: thanks to this, the system recognizes you and can recommend you job ads that match your job search criteria registered on pole-emploi.fr. By clicking on this button, you agree that our research team collects anonymized data related to your navigation on the application.

If you want to know more details, [click here](#).



 Log in with Pôle emploi




What is the purpose of this test?

This test is carried out as part of Pôle emploi's efforts to improve its assistance services. If you agree to participate, you will see personalized job recommendations. When you first log in, you may be asked to answer a few questions about your job search priorities.

If you agree to participate in this test, you will be given access to job postings recommendations. You may log back in at any time to view updated job ads recommendations.

The connection and navigation data on this application as well as all the answers you might give to any questions asked during this test will be analyzed anonymously and solely for statistical purposes by an authorized research team. They will have no impact on how your caseworker will handle your file. Some information on our research will be made public for the purposes of scientific validity, but no information that could reveal your identity will be disclosed.

 Pôle emploi met à votre disposition différents types d'affichage

[version graphique](#)

[version contrastée](#)

Figure D.6: Homepage

Rank the job ads

What do you think of the following job ads?



Sort them from those you like the most (at the top of the page) to those you like the least (at the bottom of the page). You can leave out a maximum of 2 ads.

Once the ranking is done, you can move on to the final stage of the test: viewing your personalized job ad recommendations.

1	Infirmier / Infirmière de soins généraux	24 - ANTONNE ET TRIGONANT	suivant grille FHF	35H Travail de nuit en CDI	Rank : 1
2	Infirmier / Infirmière en gériatrie (H/F)	16 - RUFFEC	Mensuel de 1852,00 Euros à 2870,00 Euros sur 12 mois	35H Horaires normaux en CDI	Rank : 2
3	Infirmier (H/F)	33 - ARES	Annuel de 28000 Euros à 35000 Euros sur 12 mois	35H Horaires normaux en CDI	Rank : 3
4	Infirmier / Infirmière de soins généraux	24 - ISSIGEAC	selon CCN 2002, profil et expérience	35H Horaires normaux en CDI	Rank : 4
5	Infirmier / Infirmière en gériatrie (H/F)	33 - STE FOY LA GRANDE	Mensuel de 2275 Euros sur 12 mois	35H Horaires normaux en CDI	Rank : 5
6	Infirmier / Infirmière de soins généraux	24 - LA TOUR BLANCHE CERCLES	Mensuel de 2200 Euros à 2400 Euros sur 12 mois	35H Horaires normaux en CDI	Rank : 6
	Infirmier / Infirmière en gériatrie	17 - FONTCOUVERTE	Mensuel de 2556 Euros	35H Horaires variables en CDI	Rank : Not ranked
	Infirmier / Infirmière en gériatrie (H/F)	33 - SOULAC SUR MER	selon grille et profil	35H Horaires normaux en CDI	Rank : Not ranked

VALIDATE

Figure D.8: Second screen

22 personalized job ads



Here are recommendations for job ads based on the search criteria of your main research, contracted during your interview with your caseworker. Click on an ad to view it in detail. Thank you for your participation in this study. You can visit this page at any time to view updated job ad recommendations.



Infirmier / Infirmière de soins généraux

KORIAN YVAN ROQUE - 24 - ISSIGEAC

Nous recherchons un/e IDE suite à une création de poste. Vous serez en charge des soins des patients en respectant le protocole sanitaire. Vous travaillez en équipe. Travail 11H / jour repos un week...

CDI
Temps plein

Publié il y a plus de 30 jours [Offre avec peu de candidats](#)



Infirmier / Infirmière en gériatrie (H/F)

SIEGE SOCIAL AUDACIA - 16 - RUFFEC

Dans le cadre d'un remplacement, au sein d'un EHPAD de 30 lits, au caractère familial, vous avez en charge les missions suivantes : - Assurer la prise en charge globale du résident afin de...

CDI
Temps plein

Publié il y a plus de 30 jours [Offre avec peu de candidats](#)



Infirmier / Infirmière de soins généraux (H/F)

ASSOCIATION LA JOIE DE VIVRE - 24 - LOLME

Assure les traitements préventifs, curatifs et palliatifs après avoir analysé, organisé et évalué les soins et les besoins de la personne. Prévient et évalue la souffrance et la dépendance des...

CDI
Temps plein

Publié il y a plus de 30 jours [Offre avec peu de candidats](#)



Infirmier / Infirmière en gériatrie (H/F)

EHPAD LE BOIS DU LORET - 33 - CENON

EHPAD de 82 résidents recrute une IDE en CDD pour des remplacements ponctuels sur des horaires 7h-16h (8h) ou 7h30-19h (10h) afin d'étoffer notre pool de remplaçants. 2.5 IDE par jour. Vous serez...

CDI
Temps plein

Publié il y a plus de 30 jours [Offre avec peu de candidats](#)



Infirmier (H/F)

CLINIQUE FONDATION WALLERSTEIN - 33 - ARES

Poste à pouvoir en contrat à durée indéterminée infirmier (e) en service chirurgie orthopédique de nuit. Le poste est à temps complet.

CDI
Temps plein

Publié il y a plus de 30 jours [Offre avec peu de candidats](#)

Figure D.9: Third screen

E Descriptive statistics

Table E.1: Comparison of full sample and participants to the pilot study

	Full sample (1)	Participants (2)
Age (in years)	38.611	44.611
Female	0.513	0.560
Any children	0.429	0.452
Married	0.386	0.470
Education: College education	0.332	0.408
Education: High school	0.238	0.209
Education: Less than high school	0.160	0.116
Education: Vocational training	0.271	0.267
Work experience (in months)	68.360	101.413
Lives in a high priority district	0.164	0.132
Receives intensive assistance from his/her caseworker	0.197	0.175
Receives unemployment benefits	0.468	0.553
First-time unemployed	0.285	0.373
Occupation targeted: Agriculture	0.032	0.025
Occupation targeted: Art and crafts	0.005	0.004
Occupation targeted: Banking, insurance, real est.	0.015	0.015
Occupation targeted: Business support services	0.158	0.247
Occupation targeted: Comm, media, digital	0.023	0.024
Occupation targeted: Construction, public works	0.060	0.040
Occupation targeted: Health	0.031	0.030
Occupation targeted: Industry	0.066	0.071
Occupation targeted: Maintenance	0.038	0.037
Occupation targeted: Performing arts	0.022	0.009
Occupation targeted: Personal services	0.186	0.185
Occupation targeted: Sales	0.157	0.153
Occupation targeted: Tourism, leisure	0.091	0.063
Occupation targeted: Transport	0.116	0.098
Monthly gross reservation wage (in euros)	1906.598	2057.865
Looking for a high-skilled white collar position	0.147	0.232
Looking for a permanent contract	0.839	0.877
Looking for a full time job	0.857	0.843
Willingness to commute: ≥ 45 min	0.255	0.286
Willingness to commute: 15 – 45 min	0.569	0.563
Willingness to commute: < 15 min	0.077	0.075
N	21063	2234

Note: A job seeker is considered as a “Participant” as soon as he/she logged in at least once to the interface. Columns (1) and (2) report mean values (as a share of the sample unless stated otherwise) for the full sample population (column 1) and for the sample of participants (column 2).

Table E.2: Summary statistics among full sample

	Full treat. (1)	Partial treat. (2)	Control (3)	P-value (4)
Age	38.308	38.319	38.387	0.375
Female	0.525	0.525	0.525	0.963
Any children	0.424	0.425	0.424	0.853
Married	0.403	0.404	0.405	0.733
Education: College education	0.330	0.331	0.330	0.924
Education: High school	0.242	0.243	0.242	0.814
Education: Less than high school	0.157	0.159	0.157	0.403
Education: Vocational training	0.271	0.267	0.271	0.154
Work experience (months)	68.159	67.558	68.244	0.275
Lives in a high priority district	0.152	0.151	0.151	0.671
Receives intensive assistance from his/her caseworker	0.188	0.186	0.187	0.577
Receives unemployment benefits	0.425	0.424	0.424	0.871
First-time unemployed	0.287	0.287	0.288	0.884
Occupation targeted: Agriculture	0.038	0.038	0.038	0.782
Occupation targeted: Art and crafts	0.007	0.008	0.008	0.433
Occupation targeted: Banking, insurance, real est.	0.015	0.015	0.014	0.286
Occupation targeted: Business support services	0.145	0.146	0.149	0.015
Occupation targeted: Comm, media, digital	0.026	0.025	0.025	0.503
Occupation targeted: Construction, public works	0.062	0.063	0.063	0.834
Occupation targeted: Health	0.037	0.035	0.036	0.243
Occupation targeted: Industry	0.071	0.070	0.071	0.973
Occupation targeted: Maintenance	0.037	0.036	0.036	0.377
Occupation targeted: Performing arts	0.022	0.021	0.021	0.854
Occupation targeted: Personal services	0.195	0.195	0.196	0.785
Occupation targeted: Sales	0.150	0.150	0.146	0.052
Occupation targeted: Tourism, leisure	0.085	0.085	0.085	0.949
Occupation targeted: Transport	0.111	0.112	0.111	0.804
Monthly gross reservation wage (in euros)	1899.180	1896.543	1899.638	0.467
Looking for a high-skilled white collar position	0.148	0.146	0.148	0.494
Looking for a permanent job	0.831	0.831	0.831	0.984
Looking for a full time job	0.855	0.854	0.857	0.174
Willingness to commute: ≥ 45 min	0.331	0.332	0.331	0.821
Willingness to commute: 15 – 45 min	0.570	0.566	0.569	0.358
Willingness to commute: < 15 min	0.099	0.101	0.099	0.414
N	83332	83332	83332	

Note: Columns (1), (2) and (3) characterize job-seekers by their treatment assignment and report mean values (as a share of the sample unless stated otherwise); column (4) reports the p-value from the F-test for joint significance of treatment coefficients in the regressions of each covariate on treatment assignment.

Table E.3: Comparison of full sample and participants to the experiment

	Full sample (1)	Participants (2)
Age (in years)	38.338	44.424
Female	0.525	0.566
Any children	0.424	0.463
Married	0.404	0.490
Education: College education	0.330	0.401
Education: High school	0.242	0.215
Education: Less than high school	0.158	0.128
Education: Vocational training	0.270	0.255
Work experience (in months)	67.987	99.879
Lives in a high priority district	0.151	0.126
Receives intensive assistance from his/her caseworker	0.187	0.172
Receives unemployment benefits	0.424	0.521
First-time unemployed	0.287	0.373
Occupation targeted: Agriculture	0.038	0.025
Occupation targeted: Art and crafts	0.008	0.007
Occupation targeted: Banking, insurance, real est.	0.014	0.018
Occupation targeted: Business support services	0.147	0.225
Occupation targeted: Comm, media, digital	0.025	0.028
Occupation targeted: Construction, public works	0.063	0.044
Occupation targeted: Health	0.036	0.033
Occupation targeted: Industry	0.071	0.071
Occupation targeted: Maintenance	0.036	0.035
Occupation targeted: Performing arts	0.022	0.013
Occupation targeted: Personal services	0.196	0.203
Occupation targeted: Sales	0.149	0.142
Occupation targeted: Tourism, leisure	0.085	0.060
Occupation targeted: Transport	0.111	0.097
Monthly gross reservation wage (in euros)	1898.454	2032.775
Looking for a high-skilled white collar position	0.147	0.226
Looking for a permanent job	0.831	0.864
Looking for a full time job	0.855	0.835
Willingness to commute: ≥ 45 min	0.332	0.346
Willingness to commute: 15 – 45 min	0.568	0.560
Willingness to commute: < 15 min	0.100	0.094
N	249996	31438

Note: A job seeker is considered as a “Participant” as soon as he/she logged in at least once to the interface. Columns (1) and (2) report mean values (as a share of the sample unless stated otherwise) for the full sample population (column 1) and for the sample of participants (column 2).

F Results from the pilot study

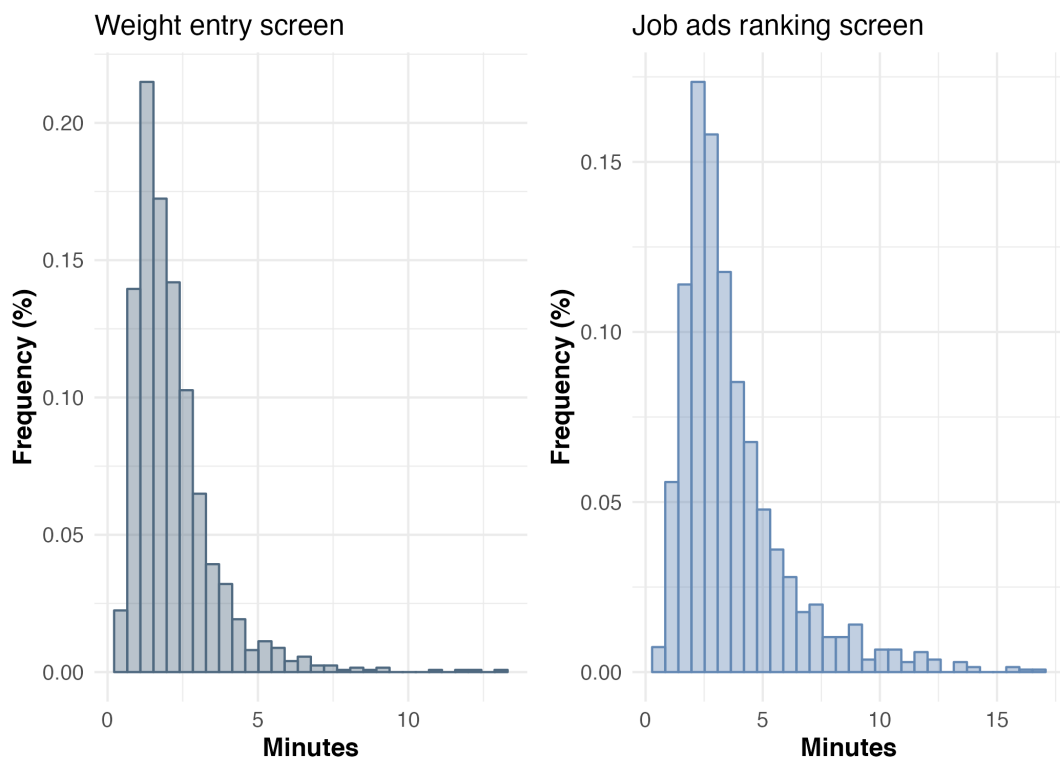


Figure F.10: Distribution of time spent on each survey step

G Analysis of stated preferences

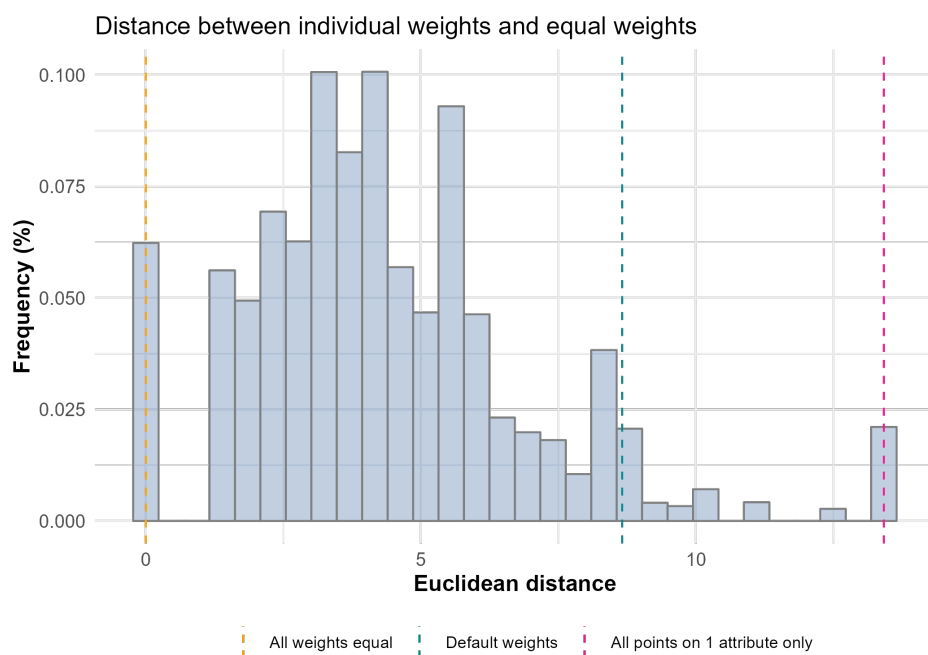
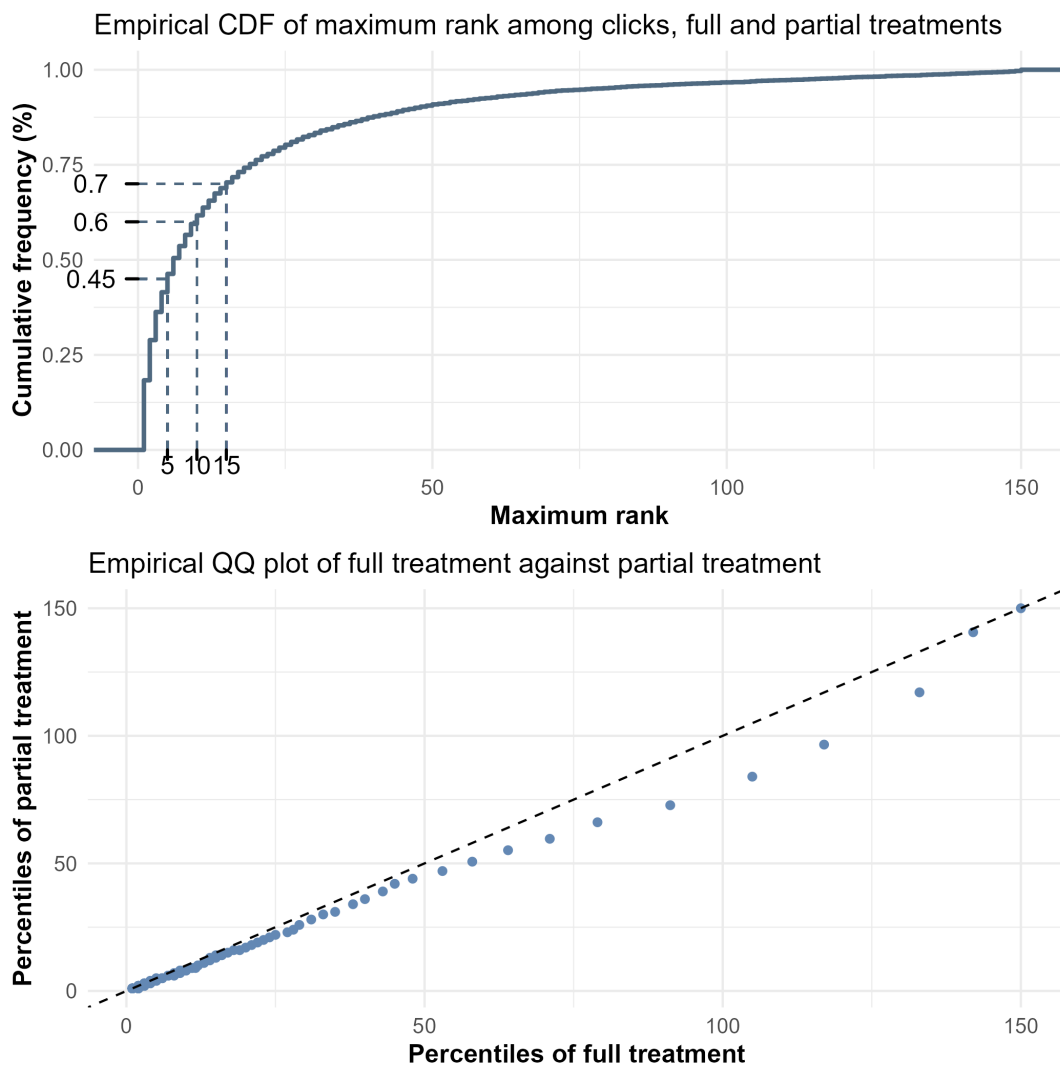


Figure G.11: Distance between individual weights and equal weights

Table G.4: Average characteristics of each cluster vs. all respondents

	Clusters				All respondents
	1	2	3	4	
Weights					
Weight assigned to: occupation	1.32	1.94	5.85	1.28	2.64
Weight assigned to: wage	1.95	2.74	2.80	6.22	3.42
Weight assigned to: distance	8.18	2.55	2.93	3.57	3.37
Weight assigned to: type of contract	1.39	3.83	1.59	1.88	2.68
Weight assigned to: working hours	2.15	3.94	1.83	2.06	2.90
Age (in years)	45.84	43.28	45.52	44.01	44.19
Female	0.61	0.55	0.55	0.55	0.56
Any children	0.48	0.48	0.42	0.46	0.46
Married	0.54	0.47	0.50	0.51	0.49
Education: College education	0.46	0.37	0.47	0.46	0.42
Education: High school	0.21	0.22	0.20	0.21	0.21
Education: Less than high school	0.13	0.13	0.10	0.11	0.12
Education: Vocational training	0.21	0.28	0.22	0.22	0.25
Work experience (in months)	100.76	91.52	111.90	107.18	100.36
Lives in a high priority district	0.09	0.14	0.11	0.10	0.12
Receives intensive assistance from his/her caseworker	0.16	0.18	0.15	0.13	0.16
Receives unemployment benefits	0.53	0.51	0.53	0.57	0.53
First-time unemployed	0.41	0.35	0.40	0.40	0.38
Occupation targeted: Agriculture	0.02	0.03	0.03	0.03	0.03
Occupation targeted: Art and crafts	0.01	0.01	0.01	0.01	0.01
Occupation targeted: Banking, insurance, real est.	0.02	0.02	0.02	0.02	0.02
Occupation targeted: Sales	0.14	0.16	0.10	0.16	0.14
Occupation targeted: Comm, media, digital	0.03	0.02	0.04	0.03	0.03
Occupation targeted: Construction, public works	0.03	0.05	0.04	0.04	0.04
Occupation targeted: Tourism, leisure	0.05	0.06	0.05	0.06	0.06
Occupation targeted: Industry	0.08	0.07	0.06	0.08	0.07
Occupation targeted: Maintenance	0.04	0.04	0.04	0.03	0.04
Occupation targeted: Health	0.02	0.03	0.04	0.03	0.03
Occupation targeted: Personal services	0.21	0.19	0.23	0.17	0.20
Occupation targeted: Performing arts	0.01	0.01	0.04	0.01	0.01
Occupation targeted: Business support services	0.25	0.22	0.24	0.23	0.23
Occupation targeted: Transport	0.08	0.11	0.08	0.09	0.10
Monthly gross reservation wage (in euros)	2,000.25	1,982.06	2,113.18	2,158.36	2,051.21
Looking for a high-skilled white collar position	0.21	0.21	0.28	0.29	0.24
Looking for a permanent job	0.83	0.91	0.85	0.83	0.87
Looking for a full time job	0.77	0.85	0.84	0.85	0.84
Willingness to commute: \geq 45 min	0.23	0.37	0.38	0.34	0.35
Willingness to commute: 15 – 45 min	0.64	0.55	0.54	0.57	0.56
Willingness to commute: $<$ 15 min	0.12	0.08	0.09	0.09	0.09
Observations	1616	8265	4050	3760	17691
Share	9 %	47 %	23 %	21 %	100 %

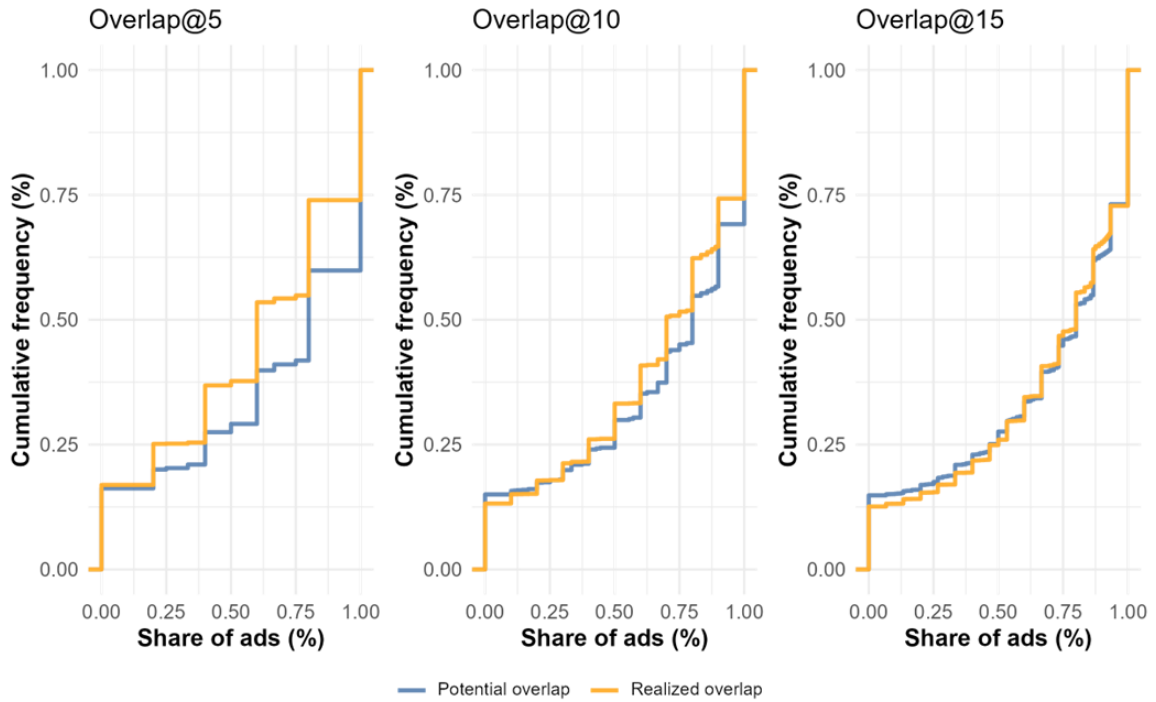
H Descriptive analysis of the recommendations



Note: The maximum rank among clicks is a variable which is measured post-treatment and which depends on the treatment: full treated job seekers received more ads than partial treated job seekers which makes the distribution of the maximum rank among full treated slightly greater than the distribution of the maximum rank among partial treated job seekers (see Q-Q plot).

Figure H.12: Maximum rank among clicks

Figure H.13: Distributions of realized and potential overlaps



Note: This set of graphs displays the cumulative distribution of job seekers (y-axis) based on the percentage overlap of job ads (x-axis) between potential recommendations and the actual ones displayed. The x-axis represents the share of job ads that overlap, expressed as a percentage, with values ranging from 0% (no overlap) to 100% (complete overlap). The blue line (“Potential overlap”) represents the overlap of job ads that could have been displayed to job seekers without the sorting issue. In contrast, the orange line (“Realized overlap”) represents the actual overlap observed in the recommendations.

I Results of the large scale experiment

I.1 Control vs. full and partial treatments: LATE estimates

While the ITT provides a useful measure of the overall effect of assignment to treatment, it does not isolate the effect of actually receiving the intended treatment. To estimate the effect among those who actually received the treatment (the compliers), we use the Local Average Treatment Effect (LATE) approach.

Although our experiment was not initially designed as an encouragement design, an unforeseen bug introduced elements of such a design into the study. The bug caused some individuals who were assigned to the treatment groups (either full or partial) to be redirected to the control group in a non-random manner. To address this issue, we employ an instrumental variable approach to obtain unbiased estimates of the treatment effect. Specifically, the variable $\text{AssignedTreatment}_i$ —which indicates whether an individual i was assigned to either the full or partial treatment group (with $\text{AssignedTreatment}_i = 1$) or the control group (with $\text{AssignedTreatment}_i = 0$)—now functions as an implicit encouragement for participants to engage with the treatment. We use $\text{AssignedTreatment}_i$ as an instrumental variable for $\text{ReceivedTreatment}_i$, a binary indicator equal to 1 if the individual actually received the treatment as intended (i.e., completed the weight entry task). By treating the assignment to the treatment groups as an encouragement, this approach allows us to isolate the causal effect of receiving the treatment on the outcome variables, despite the complications introduced by the bug.

For individuals in the treatment groups who exited the experiment early, before engaging with the treatment (not completing the weight entry task), we assign a value of 0 to all outcome variables. They are defined as $\text{AssignedTreatment}_i = 1$ and $\text{ReceivedTreatment}_i = 0$.

LATEs are estimated using Two-Stage Least Squares (2SLS) with the following specification:

$$Y_i = \alpha + \beta \text{ReceivedTreatment}_i + \varepsilon_i \quad (\text{I.3})$$

where $\text{ReceivedTreatment}_i$ is a dummy equal to one if the individual was actually treated (i.e., completed the weight entry task and received recommendations, either personalized or standard), which is instrumented by a dummy denoted $\text{AssignedTreatment}_i$ equal to one if the respondent was assigned to the treatment group. The following first-stage regression is used:

$$\text{ReceivedTreatment}_i = \gamma + \delta \text{AssignedTreatment}_i + \eta_i \quad (\text{I.4})$$

The coefficient β in Equation (I.3) indicates the effect of having received the treatment on the outcome variable Y_i .

The analysis relies on several critical conditions that are met by the design of the experiment. First, the monotonicity condition is satisfied, as there are no defiers; individuals assigned to the control group could not have completed the weight entry task because this task was not part of their treatment. Thus, if they were in the control group, they could not have defied their assignment by completing the weight entry task, which only participants in the treatment groups were asked to do. Second, the exclusion restriction holds, meaning that the assignment to the treatment groups influences the outcome exclusively through the receipt of the treatment. This ensures that no other channels affect the outcome, allowing us to isolate the effect of the treatment itself. Finally, our instrument, the assignment to the treatment groups, is highly relevant. There is a strong correlation between being assigned to a treatment group and the likelihood of receiving the treatment as intended, see table I.5.

Table I.5: First stage regressions: Probability of receiving full and partial treatments

Dependent variable	Received full treatment (1)	Received partial treatment (2)
Treatment assigned (ref. control):		
Full treatment	0.849 (0.003)***	
Partial treatment		0.846 (0.003)***
Strata fixed effects	yes	yes
N	20 832	20 995

Note: This table shows the first stage regression results from equation I.4. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level.

Full vs. control: LATE estimates The LATE estimates presented in Table I.6 illustrate the impact of receiving the full personalization treatment on job seekers' interactions with the job recommendations. The results indicate a significant negative effect across all three outcome variables: the number of clicks on recommendations, intentions to apply, and actual applications on recommendations. Specifically, receiving the full treatment resulted in a reduction of 0.236 clicks on recommendations, representing a 34.4% decrease relative to the control group's mean of 0.686 clicks. Additionally, the full treatment led to a decrease of 0.039 in intentions to apply, reflecting a 35.8% reduction compared to the

control group’s mean of 0.109 intentions. Finally, the number of actual applications submitted decreased by 0.016, amounting to a 39.0% decrease from the control group’s mean of 0.041 applications. In addition, the number of connections to the online platform also significantly decreased for the full treatment group. Specifically, receiving the full treatment led to a reduction of 0.371 connections, representing a 31.3% decrease relative to the control group’s mean of 1.186 connections.

Table I.6: Impact of receiving the full treatment

Dependent variable	Clicks on recommendations (1)	Intentions to apply on recommendations (2)	Applications on recommendations (3)	Connections to online platform (4)
Treatment status (ref. control):				
Full treatment	-0.236 (0.028)***	-0.039 (0.010)***	-0.016 (0.006)**	-0.371 (0.024)***
Strata fixed effects	yes	yes	yes	yes
N	20 832	20 832	20 832	20 832
Control mean	0.686	0.109	0.041	1.186

Note: This table shows 2SLS results from equation I.3. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

Partial vs. control: LATE estimates The LATE estimates presented in Table I.7 show how completing the weight entry task affects job seekers’ interactions with the job recommendations. The results indicate a significant negative effect on all three outcome variables: the number of clicks on recommendations, intentions to apply, and actual applications on recommendations. Specifically, receiving the partial treatment resulted in a reduction of 0.315 clicks on recommendations, representing a 45.7% decrease relative to the control group’s mean of 0.689 clicks. Additionally, the partial treatment led to a decrease of 0.057 in intentions to apply, reflecting a 52.8% reduction compared to the control group’s mean of 0.108 intentions. Finally, the number of actual applications submitted decreased by 0.023, amounting to a 54.8% decrease from the control group’s mean of 0.042 applications. In addition, the number of connections to the online platform also dropped substantially for the partial treatment group. Receiving the partial treatment resulted in a reduction of 0.395 connections, representing a 33.6% decrease relative to the control group’s mean of 1.177 connections.

I.2 Full treatment vs. partial treatment

For each outcome, we estimate the following model:

Table I.7: Impact of receiving the partial treatment

Dependent variable	Clicks on recommendations (1)	Intentions to apply on recommendations (2)	Applications on recommendations (3)	Connections to online platform (4)
Treatment status (ref. control):				
Partial treatment	-0.315 (0.027)***	-0.057 (0.009)***	-0.023 (0.006)**	-0.395 (0.024)**
Strata fixed effects	yes	yes	yes	yes
N	20 995	20 995	20 995	20 995
Control mean	0.689	0.108	0.042	1.177

Note: This table shows 2SLS results from equation I.3. Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

$$Y_i = \alpha + \beta \text{ReceivedFull}_i + \mu Z_i + \varepsilon_i \quad (\text{I.5})$$

where Y_i represents the outcome of interest for job seeker i , ReceivedFull_i is a binary variable indicating whether the job seeker received recommendations from the personalized algorithm or from the standard algorithm, and Z_i is a vector of controls accounting for the stratified randomization design.

This equation is estimated on the subsample of job seekers who actually received the treatment to which they were assigned. Specifically, job seekers who left the interface without filling in their weights (attriters) and those who were redirected to the control group due to a bug (non-compliers) were removed from the sample. The random assignment of treatments, combined with the orthogonality of attrition and the bug to treatment assignment, ensures that ReceivedFull_i is uncorrelated with ε_i . Thus, the coefficient β identifies the Average Treatment Effect (ATE) on the respondents, indicating the effect of receiving personalized recommendations. The results are presented in Table I.8.

Table I.8: Impact of the personalization treatment on recommendations appreciation and adoption

Dependent variable	Clicks on recommendations		Intentions to apply on recommendations		Applications on recommendations		Connections to online platform
	# clicks on:		# intents to apply on:		# applications on:		
	\mathcal{R}_{seen} (1.A)	$\mathcal{R}_{overlap}$ (1.B)	\mathcal{R}_{seen} (2.A)	$\mathcal{R}_{overlap}$ (2.B)	\mathcal{R}_{seen} (3.A)	$\mathcal{R}_{overlap}$ (3.B)	(4)
Treatment status (ref. partial treatment):							
Full treatment	0.084 (0.025)**	-0.012 (0.022)	0.018 (0.008)**	0.003 (0.008)	0.008 (0.004)**	0.001 (0.004)	0.014 0.012
Strata fixed effects	yes	yes	yes	yes	yes	yes	yes
N	17 678	17 678	17 678	17 678	17 678	17 678	17 678
Partial treatment mean	0.522	0.5	0.08	0.077	0.028	0.028	1.063

Note: This table shows ATE results from equation I.5, measured on the sample of respondents (those who entered their weights). Robust standard errors are reported in parenthesis. */**/** indicates statistical significance at the 10%-5%-1% level. All regressions include strata fixed effects.

Appendices References

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4), November 2010.

Chapter 3

Job Seekers' Responses to AI Job Recommendations: Insights from a Field Experiment

This chapter is based on a joint work with Guillaume Bied (University of Ghent) ¹.

Abstract: Job recommender systems promise to reduce labor market frictions by matching job seekers with suitable vacancies. However, algorithm aversion can limit or negate these potential gains. This paper measures algorithm aversion among job seekers through a field experiment. We introduce a state-of-the-art job recommender system at the French Public Employment Service (PES), combining an expert system and a machine learning model based on rich PES data, including textual data and past hires. This design allows us to manipulate perceptions of the recommendation source as either human- or algorithm-driven. Our findings show a general tendency towards algorithm aversion, with job seekers showing reduced interest in vacancies labeled as algorithm-recommended. Using machine learning methods, we identify substantial heterogeneity in this aversion: despite overall aversion, some job seekers are algorithm-friendly, preferring algorithm-labeled recommendations over human-labeled ones.

¹This chapter draws on work from the Vadore project, conducted with Philippe Caillou (UP-Saclay/LISN/INRIA), Bruno Crépon (CREST), Christophe Gaillac (University of Geneva), Solal Nathan (UP-Saclay/LISN/INRIA), and Michèle Sebag (UP-Saclay/LISN/INRIA/CNRS). It was also a partnership with France Travail, the Public Employment Service in France. We thank Paul Beurnier, H el ene Caillo, Pierre-Antoine Corre, Yann De Coster, Thierry Foltier, Cyril Nouveau, Camille Qu er e, S ebastien Robidou, and Chantal Vessereau for their operational support. We extend our gratitude to Mitia Oberti for launching the experimental phase. This research was supported by the DATAIA convergence institute as part of the "Programme d'Investissement d'Avenir" (ANR-17-CONV-0003) operated by CREST and LISN. The authors retained full intellectual freedom, and any errors are our own. We received IRB approval from the Paris School of Economics (2021-026) and pre-registered at AEA RCT registry (<https://doi.org/10.1257/rct.11157-1.2>).

Introduction

Recommender systems (RS) have become a ubiquitous feature of online platforms [Aggarwal, 2016], influencing numerous aspects of our daily lives by suggesting items tailored to our individual preferences (e.g., songs or movies on streaming services, products on e-commerce sites, and content on social media feeds). These systems address the issue of information overload by filtering vast amounts of data, providing users with relevant options, and enhancing their overall experience. In the context of job search, job recommender systems, a specific type of RS, employ AI models to match job seekers with vacancies that align with their skills, preferences, and career goals [Mashayekhi et al., 2024]. By analyzing the extensive data from online job platforms, including user behavior and vacancy characteristics, these systems aim to reduce search costs and improve job matching in the labor market.

Recognizing this, Public Employment Services (PES) in many countries are adopting AI-driven RSs to address inefficiencies in traditional job-matching processes, providing personalized, data-driven recommendations tailored to individual job seekers [World Bank, 2023]. Field experiments, such as those conducted by B elot et al. [2019, 2022], Altmann et al. [2022], show that automated occupation recommendations can expand job seekers' application scope, increase their chances of securing interviews, and improve employment metrics like hours worked and earnings. Similarly, studies focusing on personalized recommendations for companies likely to recruit in France [Behaghel et al., 2024] and on personalized job recommendations in Sweden [Le Barbanchon et al., 2023] demonstrate positive, albeit modest, effects on employment rates.

While promising, these outcomes are constrained by a considerable challenge – algorithm aversion. This phenomenon, in which users are reluctant to trust or engage with algorithmic recommendations to the same extent as they would with human-generated ones, restricts the comprehensive integration of AI tools in job search and related markets. To fully capitalize on the potential of AI-driven job recommender systems, it is essential to comprehend and address this aversion.

The literature on attitudes towards algorithms offers two competing perspectives. The concept of algorithm aversion was first introduced by Dietvorst et al. [2015], who identified that users tend to lose trust in algorithms more quickly than in humans, particularly following errors. Subsequent studies, such as those by Jussupow et al. [2020] and Mahmud et al. [2022], expanded the definition of algorithm aversion into a broader preference for human decision-making, even when no errors are present. In particular, Jussupow et al. [2020] define algorithm aversion as a biased evaluation process in which users undervalue algorithmic decisions compared to those made by human agents, who

can range from recognized experts to average individuals. This suggests that reducing the visibility of recommendations' algorithmic origins in job recommendations could mitigate aversion. Conversely, another stream of research points to algorithm appreciation, where users exhibit trust and even preference for algorithmic recommendations, especially when algorithms are perceived as objective, data-driven, and suitable for the task at hand [Jussupow et al., 2020, Mahmud et al., 2022]. This perspective implies that making users aware that recommendations are algorithmically generated could enhance their engagement with the system.

This leads us to the central question: how should algorithmic recommendations be framed to foster user engagement? To address this question, we conduct a large-scale Randomized Controlled Trial (RCT) with the French PES, employing a between-subject design to test job seekers' responses to recommendations labeled as algorithmic, human-curated, or without any source attribution. In our experiment, job seekers receive recommendations generated by a single hybrid model that integrates an expert system module—simulating human decision-making—with a state-of-the-art machine learning (ML) model trained on past hiring data (described in Bied et al. [2023]). This setup allows us to randomly vary the displayed source, presenting the recommendations as either human-driven (expert system) or algorithmically-driven (ML model) without changing their content. Participants are presented with five job ads recommended by our hybrid algorithm through an online survey designed to mimic real-world job search experiences. They are first asked to rate each recommendation using “thumbs up/thumbs down” buttons. Next, the same ads are displayed with clickable links, allowing participants to visit the PES page for each job and apply if they wish. We measure job seekers' aversion by comparing the outcomes of the algorithm group to those of the human group, in line with the literature. We focus on three key outcomes, capturing the engagement with recommended ads at different stages: initial interest, clicks for more details, and application submissions.

Our analysis reveals a clear aversion to algorithmic job recommendations. On average, job seekers exposed to algorithm-labeled recommendations express significantly less interest and engage less with the recommendations, as indicated by a lower number of clicks. We estimate that algorithm aversion reduces the total number of declared interests (“thumbs up”) and clicks by approximately 10% and 15% respectively, compared to recommendations framed as coming from human experts. While these effects are particularly evident in early-stage engagement metrics like interest and clicks, we do not observe statistically significant effects on job applications, likely due to limited statistical power at this final stage of engagement.

We further explore the role of recommendation quality in shaping algorithm aver-

sion. Quality in recommendation systems can be assessed in various ways, such as the predictive accuracy of recommended matches (i.e., whether the recommended matches will lead to actual hires) or the alignment with user preferences. We choose the latter approach because it is directly observable to job seekers, serving as a practical proxy for perceived quality from the user’s perspective. Specifically, our quality measure is a job seeker–centered matching score that evaluates how well a recommendation aligns with the job seeker’s stated preferences across multiple dimensions: occupation, contract type, location, salary, and working hours. Our findings reveal that high-quality recommendations—whether labeled as algorithmic or human-generated— elicit more engagement from job seekers. However, the aversion to algorithm-labeled recommendations persists regardless of the recommendation quality. In other words, even when the algorithm provides matches that closely align with user preferences, users remain less inclined to engage compared to when recommendations are framed as human-generated. This underscores the importance of understanding how individual job seekers characteristics influence responses to algorithmic systems and indicates that addressing algorithm aversion requires more than just improving recommendation quality.

Following the idea that algorithm aversion may be influenced more by user characteristics than by recommendation quality, we conduct a more detailed analysis of this phenomenon. Using machine learning techniques, specifically the approach developed by [Chernozhukov et al. \[2023\]](#), we identify varying responses among different subgroups of job seekers to algorithm-labeled recommendations. Our analysis reveals that attitudes towards algorithmic recommendations are diverse. Interestingly, some job seekers show a preference for algorithmic recommendations over those presented as human-curated, a behavior known as “algorithm appreciation” [[Logg et al., 2019](#)]. We identify two main groups: the 20% most “algorithm-friendly” job seekers, who respond most positively to algorithm-labeled recommendations, and the 20% most “algorithm-averse” job seekers, who show the strongest preference for recommendations framed as human-generated. For the algorithm-friendly group, the ATE of being informed that the recommendations are algorithmic rather than human is positive, leading to a 71% increase in the interest rate per job recommendation and a 46% increase in the click rate, relative to the means for the human-labeled group (0.18 for interest rate and 0.07 for click rate). In contrast, the algorithm-averse group shows a negative ATE: being told that the recommendations are algorithmic results in an approximately 103% decrease in the interest rate and a 79% decrease in the click rate compared to the human-labeled group means. On average, there are more algorithm-averse job seekers, which accounts for the overall aversion observed in our study.

By focusing on these two extreme groups, we are able to identify distinct patterns

that help to elucidate the characteristics that contribute to these opposing attitudes. In this analysis, we utilise clicks on recommendations as the primary outcome measure, as this is the most policy-relevant indicator of job seeker engagement, directly reflecting the likelihood of further exploration of job opportunities. The results indicate that younger job seekers, particularly those under the age of 35, are disproportionately represented within the algorithm-friendly group. Furthermore, this group is also characterised by a higher proportion of job seekers who are facing prolonged periods of unemployment and approaching the exhaustion of their unemployment benefits. This suggests that these individuals may have a more pressing need for effective job placements. Additionally, those who are less autonomous and who receive intensive support from the PES (who see their case worker more often) are over-represented in this group. Moreover, they are disproportionately represented in labor markets where the PES platform has a high level of market penetration. This indicates that a significant proportion of job vacancies are posted on the PES platform, in comparison to other general job platforms that are not specifically designed for registered job seekers. Furthermore, they are over-represented in tighter labor markets, where the number of job vacancies exceeds the number of job seekers. Conversely, older job seekers, particularly those over the age of 50, are more prevalent in the algorithm-averse group. This group also exhibits higher wage expectations and shorter periods of unemployment. The PES provides only moderate to minimal support to this group, and they are more likely to be in labor markets with lower PES market penetration, where relatively few job ads are posted on the PES platform in favour of other general job platforms. Additionally, they are more commonly found in competitive labor markets, where the number of job seekers surpasses available vacancies. These findings indicate that both algorithm aversion and appreciation are shaped by a combination of factors, including age, unemployment duration, labor market conditions, and the level of support provided by the PES.

From a policy perspective, our findings indicate that enhancing the quality of recommendations may not be a sufficient strategy for mitigating algorithm aversion. Instead, it is imperative that recommendations are carefully framed in order to minimize the negative effects of algorithm labeling and to maximize the benefits of these tools. Our results also demonstrate that there is no universal approach to framing recommendations. The heterogeneity in job seekers' responses indicates that the framing of the recommendations should be tailored based on individual characteristics, or alternatively, that a neutral framing should be used that allows job seekers to opt into learning more about the recommendation source if desired.

Our study contributes to several strands of the literature. First, a key contribution is our focus on algorithm aversion within the context of a high-stakes task: job recom-

recommendations for job seekers registered at a Public Employment Service (PES). Although algorithm aversion has been the subject of investigation in domains such as joke recommendations and financial forecasting, the implications are considerably more significant in the context of the labor market. This is underscored by the European Union’s AI Act, which categorizes AI systems utilized in employment as “high-risk” applications [European Commission, 2021]. The AI Act acknowledges the potential for AI-driven decisions in recruitment and job placement to have significant and enduring consequences for individuals. Indeed, research indicates that user responses to algorithms are contingent upon the nature of the task. In the context of subjective tasks, such as entertainment or tasks involving personal preferences, users tend to prefer human judgment. This is based on the assumption that algorithms are unable to adequately capture individual uniqueness or personal context [Prah and Van Swol, 2017, Yeomans et al., 2019, Longoni et al., 2019]. Conversely, algorithms are seen as more effective for objective, data-driven tasks where human-like decision-making is less necessary [Castelo et al., 2019]. In the HR field, perceptions of algorithmic versus human recruiters vary, with studies reporting mixed results regarding their perceived fairness, competence, trust, and usefulness [van Esch et al., 2021, Wesche and Sonderegger, 2021, Choung et al., 2023]. In the context of job ad recommendations, Ochmann et al. [2020] found that integrating human-like characteristics into AI job recommender systems can enhance their acceptance among job seekers. A closely related study by Bana and Boudreau [2023] investigated algorithm aversion in a university-based labor market platform, focusing on undergraduate and graduate students. Their study revealed that job seekers were less likely to pursue opportunities recommended by an algorithm compared to those recommended by a human manager, highlighting significant algorithm aversion. However, their context—students on an academic job platform—differs substantially from that of our study. The job seekers registered with the French PES represent a more diverse and vulnerable population, facing a wider array of employment challenges. Thus, while Bana and Boudreau [2023] provides valuable insights, our study extends the literature by focusing on a broader and more representative sample, exploring algorithm aversion in a real-world setting where engagement with algorithmic tools directly impacts employment outcomes.

Our second contribution lies in exploring how the quality of job recommendations—specifically, the alignment between algorithmic recommendations and job seekers’ preferences—affects algorithm aversion. Existing literature suggests that an algorithm’s performance significantly shapes user attitudes, as users often lose trust in algorithms after encountering errors, emphasizing the importance of accuracy in mitigating algorithm aversion [Dietvorst et al., 2015, Bogert et al., 2021]. However, in job recommendation systems, users typically lack visibility into the algorithm’s absolute accuracy, which is defined as its ability to

match job seekers with positions that result in actual hires. In this context, [Laumer et al. \[2018\]](#) identified “performance expectancy” (i.e., the perceived usefulness of a technology [[Venkatesh et al., 2003](#)]) as a critical factor influencing job seekers’ intentions to use these systems. Yet, their study did not investigate how job seekers’ reactions differ between algorithmic and human-generated recommendations, leaving a gap in understanding how perceived performance affects algorithm aversion. Furthermore, the literature highlights the importance of “outcome favorability”—whether the results produced by an algorithm are seen as favorable or beneficial to the user—as a key factor in user attitudes. The “outcome favorability bias” refers to the tendency of users to view a decision-making process more positively when they receive favorable outcomes [[Wang et al., 2020](#)]. In [Choung et al. \[2023\]](#)’s study, a 2x2 design was used in which both the source of the evaluation (AI algorithm or human) and its favorability (favorable or unfavorable) were randomized across participants. After reviewing a job candidate’s profile for a job opening, participants were shown an evaluation of this profile, either attributed to an AI or a human. The study found that outcome favorability bias was stronger when the evaluation was attributed to a human, i.e. participants reacted more negatively to unfavorable outcomes from a human than from an AI. However, how outcome favorability (i.e., the alignment between users’ expectations and the recommended items) affects advisory algorithms, such as job recommendation systems, remains under-explored.

The third contribution of this paper is a systematic exploration of how user characteristics shape attitudes toward algorithmic job recommendations. Using a data-driven approach free from a priori assumptions, we uncover strong heterogeneity in job seekers’ responses, revealing the simultaneous existence of both algorithm-averse and algorithm-friendly groups. This extends previous research, which has demonstrated that factors such as psychological traits, demographics, and familiarity with algorithms significantly influence attitudes toward algorithmic systems, as detailed in [Mahmud et al. \[2022\]](#). For instance, individuals with high self-esteem may feel demeaned by algorithmic evaluations, while those confident in their abilities often prefer their own judgment over algorithmic suggestions, especially in areas where they have expertise [[Lee, 2018](#), [Logg et al., 2019](#)]. Attitudes also vary by age and gender: older individuals and women, in some cases, perceive algorithmic decisions as less beneficial, though this perception is not always consistent across sectors [[Araujo et al., 2020](#), [Logg et al., 2019](#), [Thurman et al., 2019](#)]. Additionally, those with lower education levels and less comfort with numbers tend to have a lower appreciation of algorithms across different industries [[Logg et al., 2019](#)]. In the HR field, [Pethig and Kroenung \[2023\]](#) found that women are more likely to choose algorithmic over human evaluators across different hiring and career-development settings, particularly when the human evaluator is male, due to the perceived objectivity

and fairness of algorithms in mitigating gender biases.

Finally, unlike previous studies that have predominantly relied on vignette or lab environments, our approach captures the natural behaviors of job seekers interacting with real-world job recommendations within the French PES. By analyzing how algorithm aversion manifests in this public employment setting we are able to provide a more comprehensive understanding of the factors influencing the adoption and effectiveness of job recommender systems.

The paper is structured as follows. Section 1 describes the randomized field experiment conducted in collaboration with the French PES and outlines the key features of our experimental design. Section 2 discusses the data and provides summary statistics of the job seekers in our sample. Section 3 presents the main empirical results, including estimates of the average treatment effects and an exploration of heterogeneity in algorithm aversion across different job seeker groups.

1 Field experiment

The goal of our field experiment is to measure aversion to algorithmic job recommendations and identify its associated factors. Using a between-subject design, we expose job seekers to five algorithm-generated job recommendations through an online survey, randomly varying the displayed source (human, algorithmic, or neutral) to isolate its effect while controlling for recommendation quality. The following subsections describe the intervention, detailing both the algorithm and the survey design, followed by the experimental design.

1.1 Intervention

The recommendation algorithm The algorithm used to generate job recommendations was developed by us, combining an expert system with a machine-learning (ML) approach to create a highly effective job-matching tool. The expert-system component, inspired from the proprietary PES matching system, mimics human expert decision-making. It uses explicit rules to evaluate compatibility between job seeker profiles and job vacancies based on ten criteria, including job seeker characteristics (like work experience and skills) and job-specific attributes (such as occupation and location). Each criterion is quantitatively assessed through an adequacy measure, reflecting the match between the job seeker’s preferences and the recruiter’s needs. A composite matching score for each pair is calculated by summing these measures, weighted by criterion importance.

The ML component is the MUSE algorithm detailed in [Bied et al. \[2023\]](#). This is a state-

of-the-art ML model based on the very rich data available at the PES, including textual data and past hires. The MUSE function computes a matching score between job ads and job seekers by combining embeddings that model geographical, skills and general aspects of job seekers and ads. These embeddings, which result from feed forward neural networks, are trained using a triplet loss, which optimizes the relative distance between a positive job ad (a match) and a negative job ad (a non-match) from the perspective of a given job seeker.

The final algorithm selects job recommendations by balancing the input from the expert system component and the ML component, using their rankings rather than their scores. The process involves: i) creating a consideration set of job ads highly ranked by both the expert system and the ML algorithm; ii) narrowing down this set by removing the half of ads that the ML algorithm ranks lowest; iii) reordering the remaining ads by the expert system score.

By balancing both the expert system and the ML algorithm, our final algorithm ranks job recommendations according to both human expertise and ML-driven predictions. This hybrid approach ensures that we can accurately describe the job recommendations as being both driven by human expertise or artificial intelligence. Additionally, this algorithm proved to be the most effective in a beta test involving 25,000 job seekers in March 2022, achieving the highest performance scores among the tested methods (see Section 2.4 in Chapter 1).

Survey design The survey, designed and conducted online, plays a central role in our experiment. It was developed using the Qualtrics platform and consists of several stages. Screenshots of each stage of the survey are included in the appendix (see Figures A.1 to A.4).

Selected job seekers were invited by email to participate in a test of a new job referral service developed by the PES. Participants accessed the survey through a link in the email, leading to a welcome page on Qualtrics. This page briefly outlined the study, emphasizing that it was quick (approximately five minutes), personalized, anonymous, and confidential, and that responses would not be shared with caseworkers (see Appendix Figure A.1).

After agreeing to participate, job seekers viewed a page displaying five job ads (the algorithm's top-5). The source of the recommendations was indicated at the top of the page (see below, Section 1.2, for details on framing). Each job listing provided key details such as job title, company name, location, salary, work conditions, and required experience. Participants expressed their interest in each job by clicking thumbs up / thumbs down buttons labeled as "I'm interested" and "Not for me" (see Appendix Figure A.2). Rating

all five ads was mandatory to proceed.

At the bottom of the page, participants had the option to provide feedback on the job suggestions by answering several questions. These questions assessed their satisfaction with the recommendations, the extent to which they felt the recommendations met the needs of job seekers in general, and their agreement with a statement about their chances of being hired for the suggested jobs. Responses were collected using a 5-point Likert scale ranging from “Strongly disagree” to “Strongly agree” (see Appendix Figure A.3). Importantly, answering these questions was not mandatory for participants to proceed to the next page, allowing them to continue even if they chose not to provide feedback ¹.

In the final stage of the survey, participants were shown the same five job recommendations again (see Appendix Figure A.4). This time, the job ads were presented with clickable links, allowing participants to view detailed information on the PES’s website and to apply for the jobs if desired. The clicks on these links were recorded.

1.2 Experimental design

Treatment groups To measure algorithm aversion, we exogenously varied the displayed source of the job recommendations while keeping the actual source—the hybrid algorithm described above—constant. This experimental setup allowed us to examine how job seekers respond differently to recommendations that are framed as human- or algorithm-generated, while the recommendations’ generation process remains identical across all participants.

We varied the displayed source of the recommendations across three conditions, ensuring that text length for each condition was identical to avoid any unintended influence based on complexity of the statements:

- **Human Source:** Recommendations are presented as selected by labor market experts. The statement used is: “Based on insights from labor market experts, we have selected 5 job ads that might suit your profile.”
- **Algorithm Source:** Recommendations are presented as generated by an algorithm. The statement used is: “Our recommendation algorithm has selected 5 job ads that might suit your profile.”
- **Neutral Source:** Recommendations are presented without any indication of the source. The statement used is: “Here are 5 job ads that might suit your profile.”

¹While this feedback data has been collected, we have not yet had the opportunity to analyze it in detail for the current chapter.

Important choices The terminology used to describe each source plays a significant role in shaping participants' perceptions of the systems being evaluated. First, the choice of the term "recommendation algorithm" (rather than alternatives like "artificial intelligence" or "computer program") for the algorithm condition is informed by insights from [Langer et al. \[2022\]](#). This study shows that terminology significantly impacts people's perceptions of algorithmic decision-making systems. The term "algorithm" is generally associated with higher levels of machine competence and complexity, which aligns with the technical nature of our recommendation system. "Algorithm" also carries less anthropomorphism than terms like "artificial intelligence" or "AI", reducing associations with human-like cognitive abilities. Furthermore, recent media coverage significantly shapes public perceptions of AI, especially since the release of ChatGPT. [Aday \[2023\]](#) notes that, in French media, 15% of all AI mentions since 2013 occur after November 2022, with nearly 900 daily mentions in March 2023 alone. This intense spotlight often portrays AI as transformative and human-like, leading to heightened and sometimes unrealistic expectations. By choosing the term "algorithm", we aim to avoid these inflated associations, keeping the focus on the specific technical aspects of job recommendation systems rather than the over-hyped image of "AI".

In the literature on algorithm aversion, human agents typically range from recognized experts (such as physicians or financial advisors) to non-experts (such as average individuals, colleagues, or study participants). This distinction allows studies to measure preferences for human versus algorithmic decision-making across a broad spectrum of perceived expertise. In our study, we made a second important choice: we frame the human group as labor market experts to enhance the credibility of the job recommendations. Participants are more likely to trust and value advice when it is perceived to come from knowledgeable, experienced sources [[Hou and Jung, 2021](#)]. This framing ensures that any observed differences in preferences are more likely to reflect genuine perceptions of competence, rather than skepticism about the source's expertise. Additionally, this approach is shaped by practical constraints: since the algorithms in our study integrate both human expertise and machine learning, presenting the human recommendations as coming from experts is the most viable option. We acknowledge, however, that this may capture a preference for human experts rather than purely measuring algorithm aversion.

Third, to control for the influence of perceived expertise, we introduce a neutral condition. Previous research by [Hou and Jung \[2021\]](#) highlights the significant role perceived competence plays in shaping participants' choices. When humans are perceived as highly competent, participants may prefer human recommendations not because of an aversion to algorithms, but because of their trust in human expertise. The neutral condition—where no source is specified—serves as a baseline, enabling us to measure

participants' preferences without the bias introduced by emphasizing either human or algorithmic expertise. However, we acknowledge that participants may still interpret neutral recommendations in their own way, potentially associating them with algorithmic sources.

Last but not least, we chose not to randomly vary the quality of the job recommendations displayed to participants. Providing lower-quality recommendations to some participants would be unethical and counterproductive to the goals of the PES. Our priority is to ensure that all participants receive job recommendations genuinely suited to their profiles. This decision was also guided by our experimental design, which employs a real algorithm. Unlike vignette experiments that manipulate various factors in hypothetical scenarios, our study is grounded in real data and actual job recommendations. This approach maintains the authenticity of the job search experience, thereby producing more reliable and generalizable results.

Sampling and timeline The sample for the field experiment was drawn from the population of eligible job seekers registered with the PES. Job seekers were eligible if they met six criteria: (1) they were registered as unemployed, (2) they were available to start working immediately, (3) they had registered their main job search criteria, including occupation, reservation wage, commuting distance, contract type, and working hours, (4) they were at least 18 years old, (5) they lived in Metropolitan France, and (6) they had a valid email address and had agreed to receive informational emails from the PES.

Eligible job seekers were randomly assigned to one of the three treatment groups through a complete randomization process. The email campaign, which began on June 14, 2023, consisted of an initial invitation followed by three reminders for those who did not participate. Each job seeker received the same invitation email, regardless of the treatment group. In drafting the invitation email, we were mindful of the challenges posed by recurrent phishing campaigns. We carefully explained the purpose of the study and reassured participants that their personal data would be protected, while also encouraging them to participate.

2 Data and sample statistics

2.1 Data

Our research leverages a rich dataset that combines administrative records from the PES with data collected from our survey experiment. This comprehensive dataset allows us to match information at the individual level, providing detailed insights into job seekers'

backgrounds, behaviors, and interactions with job recommendations.

Pre-treatment data The administrative data from the PES offer extensive information on various aspects of job seekers and job advertisements. It contains detailed sociodemographic characteristics, including age, gender, geographical location, marital status, and number of children. Regarding job search criteria, we have data on the occupation targeted, preferred type of contract (e.g., full-time, part-time, permanent, temporary), acceptable distance from home, and reservation wage. We also possess detailed résumé information, including education level, work experience, languages spoken, and possession of a driving license. These details provide insight into the qualifications and skills of the job seekers. Additionally, we track their pre-treatment job search behavior by recording the number of connections each job seeker made to the PES's search engine, which serves as an indicator of their engagement with the PES's platform before the treatment. Information on unemployment history is also available, encompassing unemployment duration and history, reception of unemployment benefits, level of assistance from a caseworker, and cause of unemployment.

In addition to these variables, we incorporate measures that capture the conditions of the job seekers' local labor market, defined as the intersection of the occupation sought and the commuting zone of the job seeker. Specifically, we include labor market tightness, defined as the ratio of vacancies to unemployed individuals within the local labor market. This variable quantifies the level of competition for jobs and the availability of job opportunities relevant to each job seeker. We also calculate the PES market penetration, which is the ratio between the number of job ads posted on the PES's platform and the total number of job ads in the labor market for the same occupation and commuting zone. This metric reflects not only the extent to which the PES's platform covers the available job opportunities in the job seeker's local market—indicating the platform's influence and reach—but also the reliance of recruiters in this labor market on the PES for hiring. A higher PES market penetration suggests that employers heavily utilize the PES's platform for recruitment.

Recommendation characteristics and quality We assess job recommendation quality from the job seekers' perspective using matching scores computed by the expert system component of our recommendation algorithm. These scores are part of the broader expert matching system that we use and that evaluates job recommendations based on both profile fit (how well the job seeker's skills and experience align with the job requirements) and search criteria fit (how well the job ad meets the job seeker's stated preferences). In this analysis, we focus specifically on the search criteria fit, which quantifies how closely a

job ad aligns with the job seeker’s preferences across dimensions such as job type, contract type, location, wage expectations, and working hours. Each individual matching score provides a granular measure of this fit, and these scores are aggregated into a composite matching score that summarizes the overall alignment. Importantly, this composite score incorporates the weighting system used by the French PES, where certain criteria—such as occupation—are given higher importance in the matching process.

Outcome variables Our survey captures how job seekers interact with the set of job recommendations provided to them, by combining both stated outcomes with implicit behavioral data.

First, job seekers indicate their initial interest by selecting “I’m interested (thumbs up)” or “Not for me (thumbs down)” for each recommended job ad. This immediate response reflects their gut reaction and initial attraction or aversion to the recommendation source. It is essential for studying algorithm aversion, as it helps identify potential biases at the very outset of the decision-making process. Second, clicking to view more details about the job ad signals a higher level of interest and a desire to explore the opportunity further. This action represents a behavioral commitment beyond initial interest, indicating a willingness to invest time and effort to learn more about the job. Analyzing this behavior provides insights into how job seekers may overcome initial biases in favor of practical considerations when exploring job opportunities. Third, submitting an application is the ultimate form of engagement, where job seekers take concrete action toward obtaining the job. This step reflects a significant level of commitment and suggests that the job seeker perceives a strong match between their preferences and the job opportunity.

Additionally, our survey gathers comprehensive feedback from participants regarding their satisfaction with the job recommendations. Although we do not analyze this qualitative data in the present study, we plan to do so in future research.

2.2 Descriptive statistics

Balance check Tables B.1 and B.2 in Appendix display descriptive statistics of variables across treatment groups before the intervention, computed on the entire sample invited to take the survey. Only two balancing tests out of 54 are significant (at the 5% level), which confirms that the randomization was successful at balancing the treatment groups.

Participation The average participation rate in the neutral group is 20.8%. Participation in the survey means that respondents opened the survey and saw at least the first page, meaning they were exposed to the source of the recommendations (displayed at the top

of the screen) as well as the list of job recommendations².

Table 3.1 confirms that treatment assignment did not influence participation rates. This result aligns with our experimental design: all job seekers received an identical invitation email, and the user experience was consistent for every participant throughout the survey. We now restrict our analysis to *job seekers who participated to the survey*.

Table 3.1: Treatment differences in participation

	Answered the survey
Source: Human	0.001 (0.006)
Source: Algorithm	0.006 (0.006)
N. Obs.	30 000
Mean value in source: neutral	0.208

Note: The table reports treatment differences regarding having participated to the survey. The model is estimated on the full sample of job seekers that received an email invitation to participate in the experiment. Robust standard errors are reported in parentheses. *, **, ***: significance at 10%, 5% and 1%.

Description of the sample of participants We now investigate the representativeness of the participant sample along the same dimensions as in Tables B.1 and B.2. Tables B.3 and B.4 provide a comparative analysis between the full sample (30,000 observations) and the sub-sample of participants (6,308 observations).

Respondants differ from the overall population in several key aspects. Respondants are older – 50% are aged 50 and above, compared to 20% in the full sample –, and more often female (58% vs. 52% in the full sample). In terms of educational attainment, 42% of respondents hold a college diploma, which is higher than the 36% observed in the full sample. Conversely, only 7% of respondents have less than a high school diploma, compared to 9% in the full sample. The duration of unemployment spells is slightly longer for respondents, with 20% having an unemployment duration of more than two years, compared to 18% in the full sample. Regarding reservation wages, 11% of respondents have a reservation wage above twice the minimum wage, in contrast to only 7% in the full sample, suggesting that respondents might be aiming for higher-paying jobs. Intensive assistance from caseworkers is less common among respondents (19% vs. 21% in the

²Participants did not necessarily fully complete the survey. If respondents did not answer specific questions, we assigned a value of 0 to their measures of interest and click, assuming that exiting the survey indicates a lack of interest in the ads.

full sample). In terms of job search behavior, 56% of respondents used the PES's search engine in the last 6 months, compared to 42% in the full sample, indicating more active job search through the PES channel among survey participants. Regarding labor market tightness, our survey respondents are fairly representative of the full sample in terms of labor market conditions: 43% of participants in both groups are in tight labor markets, while balanced and loose markets are represented by 23% and 28% of respondents, respectively. Finally, the PES market penetration ratio reveals some differences between the groups. This ratio reflects the share of job postings on the PES platform relative to all job ads in the market, indicating the prominence of PES as a recruiter sourcing channel. In the full sample, 33% of job seekers operate in markets where the PES's job platform is a major player, while 38% are in markets where the PES's job platform is a minor player. Among respondents, the proportion in major PES markets is slightly lower at 31%, with a higher proportion in minor PES markets (42%). This suggests that respondents might be more exposed to varied job search channels beyond the PES.

Descriptive analysis of declared interest and user implicit engagement We now analyze the relationship between job seekers' declared interest in job advertisements (stated data) and their subsequent engagement, measured through clicks and applications (implicit data), pooling all respondents together. Interactions based on the source of recommendations will be examined in a later section.

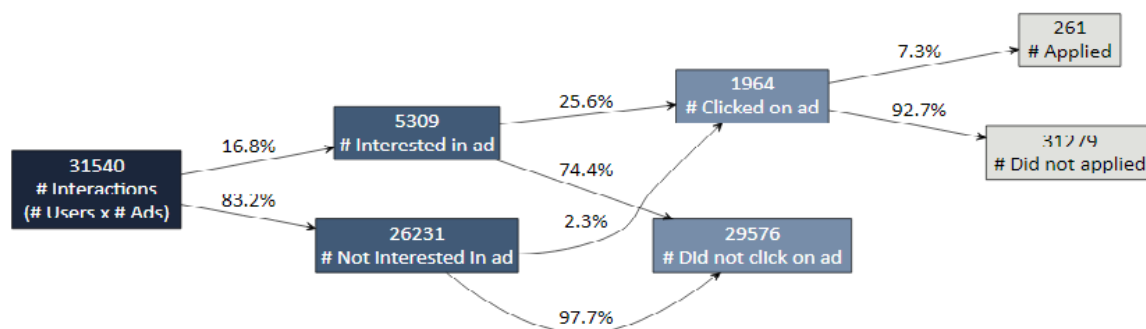
Figure 3.1 illustrates the stages of this engagement process: declaring interest, clicking on ads, and finally, submitting applications. Of the 31,540 interactions between job seekers and ads, only 16.8% resulted in declared interest. At the interaction level, 25.6% of instances where job seekers expressed interest resulted in a click on the ad, indicating a significant drop-off in engagement. This gap highlights the difference between initial curiosity and the decision to take further action. Conversely, a small percentage (2.3%) of job seekers who did not express any initial interest in ads still ended up clicking on them, suggesting that certain ads can attract attention despite an apparent lack of initial appeal. As the funnel progresses, the conversion from clicks to applications becomes even more selective. Out of the 1,964 clicks, only 7.3% resulted in applications. This further decline indicates a rigorous evaluation process on the part of job seekers; not all ads that capture attention ultimately meet their criteria for a serious application³.

These patterns suggest that declared interest is an initial but imperfect predictor of engagement. Job seekers rarely click without expressing interest, using interest as a broad

³The heatmap in Figure B.5 (in the appendix) shows the percentage per recommendation of job seekers who clicked on an ad after declaring interest. For example, 31.7% of those interested in Ad 1 clicked on it, while only 16.1% did so for Ad 5. This discrepancy is expected, as the ads displayed represent the top five results from our recommendation algorithm that ranks ads by relevance.

filter and clicks as a more refined evaluation. Applications, the most selective stage, reflect careful consideration of multiple factors. Application submission is the most selective stage, implying a careful weighing of multiple criteria.

Figure 3.1: Conversion funnel



Note: This flowchart visualizes the sequential interactions and actions taken by users in relation to job ads, starting from their initial interaction to the final application. The numbers and percentages displayed are computed on the sample of respondents from the three treatment groups.

3 Results

3.1 On average, job seekers have algorithm aversion

We estimate the Average Treatment Effects (ATE) by aggregating outcomes across all five job recommendations received by each job seeker, capturing their overall engagement with the recommendations. The analysis considers six distinct outcomes, categorized according to their nature as either continuous or binary. The continuous outcomes include the total number of times a job seeker expressed interest (i.e. clicked on “thumbs up”) in the recommended job ads (“Interest (#)”), clicked to obtain more information about the recommendations (“Click (#)”), and submitted an application on the recommended ads (“Application (#)”). The binary outcomes consist of indicators equal to 1 if the job seeker expressed interest in at least one of the recommendations (“Interest (Any)”), clicked on at least one recommendation (“Click (Any)”), or submitted at least one application (“Application (Any)”).

For each outcome, we estimate the following model using Ordinary Least Squares

(OLS):

$$Y_i = \beta_0 + \beta_1 \cdot T_i^{\text{algo}} + \beta_2 \cdot T_i^{\text{neutral}} + \varepsilon_i \quad (3.1)$$

where Y_i represents the outcome for job seeker i , such as whether they expressed interest, clicked, or applied at least once. The variables T_i^{algo} and T_i^{neutral} are binary indicators for whether job seeker i was assigned to the treatment group where recommendations were labeled as originating from an algorithm or from a neutral source, respectively. The baseline category, omitted from the model, corresponds to the case in which the source of the recommendation is attributed to human experts. We use the human source as the baseline to directly measure algorithm aversion, as defined in the literature as negative behaviors and attitudes toward algorithms compared to human agents [Jussupow et al., 2020]. Therefore, a negative coefficient β_1 would indicate algorithm aversion, as it would suggest lower engagement with algorithm-labeled recommendations. The neutral label is included in the regression to gain insights into how job seekers perceive recommendations when the source is not explicitly stated. This approach is designed to avoid influencing their perception: when no source is indicated, job seekers may interpret the recommendations as originating from an algorithm, from experts, or may not consider the source at all. This model is estimated using the subset of job seekers who participated to the survey⁴.

Alternative specifications are also considered to ensure the results' robustness. A negative binomial model, which is well-suited to handle the count nature of the data, is considered in Appendix C.1. A model estimated at the job seeker-vacancy pair level is also detailed in the following section (to further inspect heterogeneity). Both specifications yield results that are consistent with our main findings, confirming the robustness of our estimates.

Table 3.2 presents the results. The findings reveal a statistically significant aversion to recommendations labeled as coming from an algorithm. Specifically, job seekers exposed to the algorithm label exhibited a 10.6 % decrease in the total number of times they expressed interest compared to the expert group, corresponding to a reduction of 0.094 points ($p < 0.05$). Furthermore, the probability of expressing interest for at least one job ad was 5.9 % lower (-0.025, $p < 0.10$) for those in the algorithm-labeled group.

This aversion extends to engagement metrics involving clicks. The total number of clicks on recommendations labeled as coming from an algorithm decreased by 13.9 % (-0.047, $p < 0.05$) relative to the expert-labeled group. Moreover, the likelihood of clicking on at least one recommendation was 11.5 % lower (-0.027, $p < 0.05$) for the algorithm-labeled group compared to the expert group. These results indicate that the negative

⁴If respondents did not fill in all the questions, we assigned a value of 0 to their measures of interest and click, assuming that exiting the survey indicates a lack of interest in the ads.

Table 3.2: Impact of perceived source on the appreciation of recommendations, experts as reference

Dependent variable	Interest (#)	Interest (Any)	Click (#)	Click (Any)	Application (#)	Application (Any)
Source: Algorithm	-0.094 (0.038)**	-0.025 (0.015)*	-0.047 (0.022)**	-0.027 (0.013)**	0.008 (0.007)	0.006 (0.006)
Source: Neutral	-0.044 (0.039)	-0.009 (0.015)	-0.033 (0.022)	-0.017 (0.013)	-0.0003 (0.007)	-0.004 (0.006)
N. Obs.	6308	6308	6308	6308	6308	6308
Mean value in experts group	0.888	0.422	0.338	0.235	0.039	0.035
P-value of $\beta_1 = \beta_2$	0.371	0.553	0.780	0.706	0.485	0.199

Note: Source: Experts is used as reference. *, **, ***: significance at the 10%, 5%, and 1% levels. Robust standard errors in parenthesis.

impact of the algorithmic label goes beyond mere interest, affecting deeper levels of engagement with the job recommendations. In contrast, differences in application behavior are minimal and not statistically significant. Neither the total number of applications nor the likelihood of submitting at least one application varies significantly between the algorithm-labeled and expert-labeled groups. However, it is worth noting that we are under-powered to detect small effects on applications, which may explain the absence of significant differences in this outcome.

Interestingly, the absence of a source label results in an intermediate level of appreciation for the recommendations, falling between the explicitly labeled “algorithm” and “expert” signals. The differences are not statistically significant when compared to expert label nor to the algorithm label⁵. These results imply that not providing a source label does not significantly shift job seekers’ engagement with the recommendations.

Overall, these findings point to a nuanced form of algorithm aversion. The significant reductions in both interest and click rates for algorithm-labeled recommendations compared to human-labeled recommendations indicate a clear aversion to algorithmic sources, as defined in the literature—i.e., as a preference for human input over algorithms [Jussupow et al., 2020, Mahmud et al., 2022]. In addition, the absence of significant differences in engagement for the neutral label suggests that this aversion is not driven by a strong preference for expert recommendations. Instead, it appears that job seekers are reacting to the specific connotations associated with the algorithmic label, rather than expressing a clear preference for human expertise.

⁵The p-values from testing $\beta_1 = \beta_2$ for each outcome are as follows: 0.3706 for the total number of declared interests, 0.5534 for declaring at least one interest, 0.7804 for the total number of clicks, 0.7061 for at least one click, 0.4846 for the total number of applications, and 0.1999 for submitting at least one application.

3.2 Exploring heterogeneous effects by recommendation quality

When the actual quality of a recommendation differs from what an individual expects, the impact may vary depending on whether the recommendation is perceived as coming from an algorithm or a human advisor. In this section, we explore whether algorithm aversion changes with the quality of recommendations, specifically in terms of how well they match job seekers' predefined search criteria. Our analysis is carried out at the pair level, with each observation representing a unique job seeker–vacancy pairing.

We begin by estimating a baseline model without interaction terms:

$$Y_{ij} = \beta_0 + \sum_{k \in \{algo, neutral\}} \beta_k \cdot T_i^k + \varepsilon_{ij} \quad (3.2)$$

In this model, Y_{ij} represents job seeker i 's engagement with vacancy j , measured by declared interest, clicks, or applications. The treatment indicators T_i^{algo} and $T_i^{neutral}$ denote whether the recommendation was attributed to an algorithm or presented neutrally, with recommendations from human experts as the baseline.

To assess treatment effect heterogeneity by recommendation quality, we extend the model by incorporating the preference matching score, which measures how well the job recommendation aligns with the job seeker's predefined search criteria. The model is specified as:

$$Y_{ij} = \beta_0 + \sum_{k \in \{algo, neutral\}} \beta_k \cdot T_i^k + \gamma \cdot P_{ij} + \sum_{k \in \{algo, neutral\}} \theta_k \cdot (T_i^k \times P_{ij}) + \varepsilon_{ij} \quad (3.3)$$

Here, P_{ij} is the preference matching score, reflecting how well the job matches the job seeker's preferences in terms of occupation, location, job type, contract terms, and wage. This variable is centered and scaled between 0 and 1. The interaction terms between the treatment indicators and the preference matching score allow us to test whether the effect of the recommendation source varies based on this dimension of recommendation quality. Both models are estimated using the sample of job seekers who completed the full survey, and standard errors are clustered at the individual level to account for within-person correlation across multiple recommendations.

The results of the heterogeneous treatment effects analysis are presented in Table 3.3. The table reports results from the two models: the baseline model (columns 1, 3, and 5) estimated from equation 3.2, and the model that includes interaction terms between the recommendation source and the preference matching score (columns 2, 4, and 6), estimated from equation 3.3.

In the baseline models (1, 3, and 5), which do not account for the quality of recommendations, we observe a negative and significant effect of algorithmic recommendations on interest and clicks, while no significant effect is found for applications. The negative coefficient for the algorithmic source suggests that, absent any consideration of recommendation quality, job seekers engage less with recommendations attributed to algorithms compared to those attributed to human experts. For neutral recommendations, there is no significant difference in engagement across all three outcomes, indicating that the neutral framing of recommendations does not significantly alter job seekers' behavior relative to expert-attributed recommendations. These findings align with the previous specification.

The results from the inclusion of the preference matching score in models (columns 2, 4, and 6) reveal several notable patterns across the different engagement outcomes. The preference matching score exhibits a strong positive association with engagement, as evidenced in model (2), which is estimated on the interest outcome. The coefficient on the preference matching score (γ in Equation 3.2) is 0.132 and statistically significant at the 1% level, indicating that job recommendations better aligned with job seekers' preferences generate a higher level of interest. This positive association holds consistently across all outcomes (interest, clicks, and applications), highlighting that job seekers place considerable value on recommendations that closely match their search criteria. The significance and magnitude of these coefficients support the validity of the preference matching score as a credible measure of recommendation quality.

The interaction terms between the treatment indicators and the preference matching score provide further insights into the role of perceived source on the relationship between recommendation quality and engagement. Specifically, the interaction between the algorithm treatment and the preference matching score (θ_{algo} in Equation 3.3) yields a slightly positive coefficient in models (2) and (6), which are estimated on the interest and application outcomes, respectively. However, the coefficient is slightly negative in model (4), which is estimated on the click outcome. Importantly, these coefficients are not statistically significant across any of the three outcomes, suggesting that the extent to which recommendations align with job seekers' preferences does not substantially influence their engagement with recommendations perceived as algorithm-generated versus human-generated.

In contrast, the coefficient associated with the interaction between the neutral treatment and the preference matching score ($\theta_{neutral}$ in Equation 3.3) is slightly negative in model (2) on the interest outcome and slightly positive in models (4) and (6) on the click and application outcomes. While these coefficients are not significant for the interest and click outcomes, they reach statistical significance for the application outcome. However, this finding is somewhat puzzling and may, in part, reflect noise rather than a genuine

effect, particularly given the limited power to detect effects on applications due to the extremely low average application rate (less than 1% per job ad).

Table 3.3: Heterogeneous treatment effects by recommendation quality

Dependent variable	Interest (1)	Interest (2)	Click (3)	Click (4)	Application (5)	Application (6)
(Intercept)	0.178 (0.006)***	0.177 (0.005)***	0.068 (0.003)***	0.067 (0.003)***	0.008 (0.0009)***	0.008 (0.0009)***
Source: Algorithm	-0.019 (0.008)**	-0.019 (0.008)**	-0.009 (0.004)**	-0.009 (0.004)**	0.002 (0.001)	0.002 (0.001)
Source: Neutral	-0.009 (0.008)	-0.007 (0.008)	-0.007 (0.004)	-0.006 (0.004)	-0.00006 (0.001)	0.0001 (0.001)
Preference matching score		0.132 (0.017)**		0.035 (0.009)***		0.009 (0.003)**
Source: Algorithm × Preference matching score		0.005 (0.023)		-0.004 (0.013)		0.006 (0.005)
Source: Neutral × Preference matching score		-0.004 (0.024)		0.014 (0.013)		0.010 (0.005)**
N. Clusters	6308	6308	6308	6308	6308	6308
N. Obs.	31 540	31 540	31 540	31 540	31 540	31 540

Note: Standard errors are clustered at the individual level. Models (1), (3) and (5) are estimated from equation 3.2. The preference matching score is centered and scaled between 0 and 1. Models (2), (4) and (6) are estimated from equation 3.3. *, **, ***: significance at the 10%, 5%, and 1% levels.

These findings suggest that algorithm aversion is driven less by the intrinsic quality of the recommendations and more by how individuals perceive the source. While our initial exploration focused on heterogeneity in recommendation quality, identifying treatment effect heterogeneity across high dimensional covariates necessitates a more systematic approach, a challenge that we address by leveraging machine learning methods in the next section.

3.3 Systematically identifying algorithm-averse and algorithm-friendly job seekers: a machine learning approach

In this section, we shift our focus from estimating the ATE to the Conditional Average Treatment Effect (CATE). The ATE captures the overall effect of algorithmic recommendations but fails to account for individual differences in responses. Job seekers may vary in their attitudes toward algorithmic suggestions: some may prefer them (algorithm appreciation), while others resist (algorithm aversion). Thus, estimating the CATE helps us understand how different subgroups, defined by their covariates X , respond to these recommendations.

The CATE is defined as:

$$\tau(X) := E[Y_1|X] - E[Y_0|X], \quad (3.4)$$

where Y_1 and Y_0 are the potential outcomes under treatment and non-treatment, and X is a vector of pre-treatment variables. By accounting for this heterogeneity, we can develop ways of presenting and describing job recommendations that better accommodate diverse preferences. Furthermore, recognizing these differences is essential for ensuring fairness: ignoring them may lead to systems that disproportionately disadvantage certain groups, thereby exacerbating existing inequalities and reducing inclusiveness.

A common approach to studying heterogeneity is to use standard regression models, interacting pre-treatment variables with the treatment indicator or stratifying data into subgroups, and testing whether the treatment effect is constant across all subgroups. These approaches have limitations. They require researchers to specify the subgroups of interest in advance, potentially overlooking important heterogeneities or leading to small, imprecise estimates for some groups. Additionally, conducting multiple hypothesis tests increases the false discovery rate (e.g. with 50 tests at the 5% significance level, the probability of incorrectly rejecting at least one null hypothesis could rise to 92%). While corrections for multiple testing exist, they still require pre-specifying potential sources of heterogeneity. Pre-analysis plans can help, but as noted by [Olken \[2015\]](#), they are costly and inflexible, often discarding valuable data.

To overcome these challenges, we rely on machine learning (ML) methods, which offer a systematic way to uncover heterogeneity without the need for predefined hypotheses. ML is particularly effective at identifying complex, hidden patterns in the data, making it a powerful tool for detecting treatment effect variations that traditional techniques may miss.

We explore the heterogeneity of algorithm aversion by examining a binary treatment: algorithm-stated source versus human-stated source. To do so, we restrict our sample to jobseekers assigned to either the algorithmic source (treatment group) or the expert source (control group). Our analysis focuses on two binary outcomes: the interest rate and the click rate for job recommendations. The dataset is structured in a long format, where each individual receives five recommendations, allowing us to incorporate more granular job-level characteristics. This structure enhances statistical power and facilitates the use of disaggregated variables specific to each job ad in our analysis.

3.3.1 Overview of the Generic Machine Learning framework

Estimating the full Conditional Average Treatment Effect (CATE) function with ML methods typically requires strong assumptions or large amounts of data, making consistent and reliable inference difficult. To address this, we follow the Generic Machine Learning (GML) approach introduced by [Chernozhukov et al. \[2023\]](#). Instead of attempting to perfectly estimate the CATE, we use ML estimators as proxies to infer key features of

treatment effect heterogeneity. By focusing on these coarser features instead of the entire CATE function, valid inference and valuable insights can be obtained without assuming the consistency or unbiasedness of the underlying machine learning proxies.

The main idea is to estimate the CATE proxy, denoted $S(X)$, using ML methods and post-process it to make inferences about important features of the CATE, such as:

1. The Best Linear Predictor (BLP) of the CATE $\tau(X)$ using the ML proxy $S(X)$.
2. The Sorted Group Average Treatment Effect (GATES): we partition individuals into groups based on $S(X)$ and estimate the average treatment effect for each group.
3. The average characteristics of the most and least affected groups, analyzed through Classification ANalysis (CLAN).

Learning ML proxies Following Chernozhukov et al. [2023], we randomly split our data into a main sample (M) and an auxiliary sample (A). Given the long format of our data, where each individual–job recommendation pair is an observation, we split the data at the individual level to maintain independence between M and A .

We train ML models on the auxiliary sample A to estimate proxies for two key quantities: (1) $B(X)$, a proxy for the baseline effect $b(X) = E[Y_0 | X]$, representing the expected outcome for untreated observations, and (2) $S(X)$, a proxy for the CATE $\tau(X)$, as defined in Equation 3.4. Here, X represents the covariates associated with each individual and job recommendation.

After training, we apply these estimates to the main sample M to ensure out-of-sample validity and prevent overfitting. To estimate the CATE, we adopt the common approach of training separate ML models for the treated and control groups using observations from the auxiliary sample. The expected outcome in the absence of treatment is estimated using the control observations, while the expected outcome under treatment is estimated using the treated observations. These trained models are then applied to the main sample to predict the baseline effect for each individual–job pair and compute the CATE as the difference between the predicted outcomes for the treated and control groups.

Testing for treatment effect heterogeneity Having obtained the estimates $B(X)$ and $S(X)$, we first estimate the following weighted linear regression model on the main sample for each of the two outcomes (interest rate and click rate):

$$Y_{ij} = \alpha' Z_{ij} + \beta_1(T_i - p) + \beta_2(T_i - p)(S_{ij} - E[S_{ij}]) + \varepsilon_{ij}, \quad (3.5)$$

with weights $w_{ij} := (p(1-p))^{-1}$. Here, Y_{ij} is the outcome of interest for individual i and recommendation j , T_i is the treatment indicator for individual i (equaling 1 if the individual is in the algorithm-label condition), and p is the known propensity score, which is constant in our pure RCT case. $S_{ij} := S(X_{ij})$, and $E[S_{ij}]$ is the mean of S_{ij} in the main sample. The matrix Z_{ij} is defined as $Z_{ij} := [1, B(X_{ij})]$. Given the multiple observations per individual, we cluster standard errors at the individual level to account for within-individual correlation.

Chernozhukov et al. [2023] show that the coefficients $(\hat{\beta}_1, \hat{\beta}_2)$ estimated from regression (3.5) approximate the Best Linear Predictor (BLP) of $\tau(X)$ using $S(X)$. More formally, they target the estimands of interest β_1 and β_2 that solve the optimization problem:

$$(\beta_1, \beta_2) \in \operatorname{argmin}_{(b_1, b_2) \in \mathbb{R}^2} E[\tau(X) - b_1 - b_2 S(X)]^2, \quad (3.6)$$

In particular, β_1 represents the ATE, and $\beta_2 = \frac{\operatorname{Cov}(\tau(X), S(X))}{\operatorname{Var}(S(X))}$ reflects how much $S(X)$ explains the heterogeneity in the treatment effects. If $S(X)$ perfectly predicts $\tau(X)$, then $\beta_2 = 1$. Conversely, $\beta_2 = 0$ indicates that $S(X)$ contains no information about $\tau(X)$. Importantly, testing whether $\beta_2 = 0$ allows for a formal test of treatment effect heterogeneity. If $\tau(X)$ is constant (i.e., no heterogeneity), then $\beta_2 = 0$, simplifying the BLP to $\operatorname{BLP}[\tau(X) | S(X)] = \beta_1$. Thus, rejecting the null hypothesis $\beta_2 = 0$ provides evidence of heterogeneity and confirms that $S(X)$ is a relevant predictor of this heterogeneity.

Further analysis of heterogeneity: identifying groups and their characteristics Once we establish the presence of heterogeneity in the treatment effect, our next step is to estimate the Sorted Group Average Treatment Effects (GATES) to identify and quantify how different groups respond to the treatment. Given the long-format structure of our dataset, where each individual i receives multiple (5) job recommendations j , we first compute the mean of the ML proxy $S(X_{ij})$ for each individual:

$$\bar{S}(X_i) = \frac{1}{5} \sum_{j=1}^5 S(X_{ij}) \quad (3.7)$$

We then divide the sample into quintiles based on $\bar{S}(X_i)$, ensuring that all observations for a given individual fall within the same quintile, which enhances the interpretability of the results.

Then, we estimate the following weighted linear regression on the main sample for

each of the two outcomes (interest rate and click rate):

$$Y_{ij} = \alpha' Z_{ij} + \sum_{k=1}^5 \gamma_k (T_i - p) \cdot \mathbb{1}\{\bar{S}_i \in G_k\} + v_{ij} \quad (3.8)$$

with weights $w_{ij} := (p(1-p))^{-1}$. Here, Y_{ij} is the outcome for individual i and job recommendation j , T_i is the treatment indicator for individual i (equaling 1 if the individual is in the algorithm-label condition), and p is the known propensity score, which is constant in our pure RCT case. $\bar{S}_i := \bar{S}(X_i)$, and G_k is the set of observations in the main sample that fall into the k th quintile of \bar{S}_i . The covariate vector Z_{ij} includes individual and job-level characteristics and is defined as $Z_{ij} := [1, B(X_{ij})]$. We cluster standard errors at the individual level to account for within-individual correlation.

Chernozhukov et al. [2023] show that the coefficients $\gamma = (\gamma_k)_{k=1}^5$, estimated from regression (3.8), capture the average treatment effects within each group G_k :

$$\gamma = (\gamma_k)_{k=1}^5 = (E[\tau(X) | G_k])_{k=1}^5 \quad (3.9)$$

Once the groups are formed, we examine the characteristics $g(X_{ij})$ of the observations in the lowest (G_1) and highest (G_5) treatment effect groups to identify the types of individuals and job recommendations most and least affected by the treatment. These characteristics are summarized as:

$$\delta_1 = E[g(X) | G_1], \quad \delta_5 = E[g(X) | G_5]. \quad (3.10)$$

We then quantify the difference between these subgroups using a simple difference-in-means t-test.

Analyzing cross-outcome treatment effect heterogeneity We also aim to explore the relationship between the heterogeneity in terms of interest and the heterogeneity in terms of clicks. To carry out this analysis, we proceed as follows. For each data split (into auxiliary and main samples), we train a ML proxy on the auxiliary sample to estimate the individual treatment effect for the interest outcome. Next, we compute the predicted treatment effect in terms of interest for each observation (i, j) in the main sample, denoted as S_{ij}^{int} . We then classify observations into quintiles based on these predicted treatment effects, in the same way as before.

After classifying observations, we estimate the following weighted linear regression

model on the main sample for the click outcome:

$$Y_{ij}^{\text{click}} = \alpha' Z_{ij} + \sum_{k=1}^5 \gamma_k (T_i - p) \cdot \mathbb{1}\{\bar{S}_i^{\text{int}} \in G_k^{\text{int}}\} + \epsilon_{ij} \quad (3.11)$$

where Y_{ij}^{click} is the click indicator for individual i and job recommendation j , \bar{S}_i^{int} represents the predicted treatment effect on the interest rate for observation (i, j) , and G_k^{int} is the set of observations in the main sample falling within the k -th quintile of S_{ij}^{int} . The weights $w_{ij} := (p(1-p))^{-1}$ ensure the correct adjustment for the known propensity score p . Notably, [Bryan et al. \[2021\]](#) refer to this method as the Conditional GATES.

Inference on the key features of the CATE: BLP, GATES, and CLAN We follow the approach from [Chernozhukov et al. \[2023\]](#) to obtain valid inference for the BLP, the GATES, and the CLAN parameters. We repeatedly split the dataset into an auxiliary sample (A) and a main sample (M), performing 100 random partitions. For each split, we compute the key parameters mentioned and report their associated confidence intervals, standard errors, and p-values. Two sources of uncertainty must be accounted for: (1) conditional uncertainty, which arises from the estimation process given a particular split, and (2) variational uncertainty, which stems from the randomness introduced by data partitioning, as different splits yield different estimates.

To address these uncertainties, we report the median of the estimates across all splits to obtain a robust point estimate for each parameter. This median-based estimator mitigates the variability that can occur from any single split. Second, to construct confidence intervals that account for variability from the random splits, we report intervals based on the medians of the upper and lower bounds of the confidence intervals across splits, ensuring the additional uncertainty from data partitioning is properly reflected. The resulting $(1 - \alpha)$ confidence interval covers the true median of the parameter across partitions approximately $(1 - 2\alpha)\%$ of the time, thus providing a conservative adjustment that accounts for the randomness introduced by the splits. For hypothesis testing, similar adjustments are made to p-values. In testing the null hypothesis H_0 against an alternative H_1 , the p-values from each split vary due to the randomness of data partitioning. A test is considered significant at the level α if at least 50% of the splits yield p-values smaller than half the significance level $\alpha/2$. Once the median p-value is determined, the final adjusted p-value is obtained by doubling it. We report these sample-splitting adjusted p-values, that control for the additional variability due to the random splits.

3.3.2 Implementation details

To learn ML proxies, we use a set of covariates defined both at the jobseeker level and the pair level, which are listed in appendix C.2.

Four machine learning algorithms—OLS, Lasso, Random Forest, and LightGBM—are considered to learn the ML proxies. OLS serves as a simple baseline to compare with more complex methods. Lasso [Tibshirani, 1996] is useful for selecting the most important variables and reducing overfitting, especially in cases with many covariates, by shrinking coefficients associated with less relevant ones to zero. Random Forest [Breiman, 2001], which combines multiple decision trees, is effective in capturing non-linear relationships and remains robust to overfitting. Finally, LightGBM [Ke et al., 2017] was selected due to its computational efficiency and strong predictive performance. LightGBM uses gradient boosting, a method that sequentially builds models by combining weak learners, typically decision trees, and where each new tree seeks to correct the errors of the previous ones. This approach captures complex, non-linear relationships while remaining robust against overfitting.

The results are based on 100 different random partitions of the data, with 50% allocated to the auxiliary sample (where the algorithms are trained) and 50% to the main sample (where the treatment effect is predicted using the ML proxies obtained from the auxiliary sample). Each ML algorithm’s parameters were tuned for every random split of the data using 5-fold cross-validation, which involves training the model on 4/5 of the auxiliary sample and selecting the parameters that yield the best performance on the held-out portion. Given that each ML method has several parameters to optimize, we tested 50 unique combinations of randomly sampled parameters and selected the combination that provided the lowest out-of-sample error. This randomized search approach has proven more advantageous compared to fixing values in the hyper-parameters search space [Bergstra and Bengio, 2012]. We used the Python `scikit-learn` package to tune and train these models on the auxiliary samples. Specifications on the main samples were estimated using R.

In order to choose a ML algorithm among the four tested ones, we use the goodness-of-fit metrics introduced by Chernozhukov et al. [2023]. The first metric, targeting the BLP, is defined as $\Lambda_{BLP} := \left| \hat{\beta}_2 \right|^2 V(\widehat{S(X)})$, where β_2 is defined in equation 3.5. The second metric, targeting the GATES, is defined as $\Lambda_{GATES} := E \left(\sum_{k=1}^5 \hat{\gamma}_k \mathbb{1}(S \in I_k) \right)^2$, with γ_k defined in equation 3.8.

3.3.3 Results on the magnitude of the heterogeneity

Table 3.4 compares the ML methods in terms of how well they estimate the BLP (Λ_{BLP}) and how well they estimate the GATES (Λ_{GATES}). We find that the Light GBM outperforms the other ML methods for both metrics, since both Λ_{BLP} and Λ_{GATES} are at their maximum for this method.

Table 3.4: Comparison of ML methods

		OLS	Lasso	Random Forest	Light GBM
Interest rate	Best BLP (Λ_{BLP})	0.0345	0.0169	0.0945	0.1045
	Best GATES (Λ_{GATES})	0.0018	0.001	0.011	0.0113
Click rate	Best BLP (Λ_{BLP})	0.0127	0.0086	0.0365	0.0427
	Best GATES (Λ_{GATES})	0.0003	0.0003	0.0008	0.0009

Note: Medians over 100 splits in half. A bold font indicate highest median values.

Table 3.5 presents the results of the BLP analysis, estimated from equation 3.5 using the ML proxies learned from the Light GBM and from the Random Forest. Providing results from Random Forest allows for a robustness check: while LightGBM performs better overall, Random Forest also produces strong results in terms of BLP and GATES estimation. Considering both algorithms enables us to demonstrate that the findings are not driven by a single algorithmic approach but hold across different machine learning frameworks. We report estimates of the ATE (β_1) and of the heterogeneity loading parameter (HET, β_2) for the two outcomes of interest: the interest rate and click rate per job recommendation. The ATE for both interest and click rates is negative, showing algorithm aversion on average.

More importantly, the heterogeneity loading parameter is large and significant for both interest and click rates: 1.239 for interest rate and 0.909 for click rate, with both adjusted p-values being less than 0.001. These results indicate substantial heterogeneity in treatment effects across individuals, suggesting that attitudes toward algorithmic recommendations vary widely. The significant heterogeneity implies that while some job seekers exhibit strong aversion to algorithmic recommendations, others may be indifferent or even favor algorithmic suggestions.

3.3.4 Group Average Treatment Effects results

We now further investigate this heterogeneity by estimating the GATES parameters. To define the groups, we divided observations into $K = 5$ groups based on the quintiles of the ML proxy predictor $S(X)$ as learned from the Light GBM. Figure 3.2 displays the

Table 3.5: Best Linear Predictor

	Light GBM		Random Forest	
	Interest rate	Click rate	Interest rate	Click rate
ATE (β_1)	-0.023	-0.008	-0.020	-0.007
90% CB	[-0.039, -0.007]	[-0.018, 0.002]	[-0.037, -0.004]	[-0.018, 0.004]
Adj. p-val.	0.009**	0.216	0.025**	0.336
HET (β_2)	1.239	0.909	0.791	0.716
90% CB	[1.009, 1.471]	[0.576, 1.228]	[0.601, 0.985]	[0.387, 1.063]
Adj. p-val.	<0.001***	<0.001***	<0.001***	<0.001***

Note: Medians over 100 splits in half. *, **, ***: significance at 10%, 5% and 1%.

estimated GATES coefficients from Equation 3.8, showing the coefficients for each group along with their confidence intervals. The figure also includes the estimated ATE for comparison. Table 3.6 presents the results of the hypothesis test assessing whether the difference in ATE between the groups with the highest and lowest treatment effects is statistically significant.

For the interest rate, the bottom 20% of job seekers show a strong negative treatment effect of -0.185 (confidence interval: [-0.225, -0.144]), with a highly significant adjusted p-value (<0.001), indicating aversion to algorithmic recommendations. In contrast, the top 20% display a positive effect of 0.128 (confidence interval: [0.088, 0.167]), also significant ($p < 0.001$). The difference between these groups is 0.307, with an adjusted p-value below 0.001, highlighting substantial heterogeneity in response to the treatment. For the click rate, the bottom 20% show a treatment effect of -0.055 (confidence interval: [-0.083, -0.028]), again statistically significant ($p < 0.001$), whereas the top 20% exhibit a slightly positive but significant effect of 0.032 (confidence interval: [0.006, 0.057]). The top-bottom difference is 0.087, with a significant p-value (<0.001), pointing to significant variation in click rate responses.

We label the bottom 20% of individuals as “algorithm-averse” because their reactions to algorithmic recommendations are significantly negative, as evidenced by the large negative treatment effects on both interest rate and click rate. This group exhibits a strong algorithm version, i.e. a strong preference for human recommendations compared to algorithmic recommendations. On the other hand, the top 20% are labeled “algorithm-friendly” because their responses to algorithmic recommendations are positive, indicating algorithm appreciation.

To determine whether individuals who exhibit higher aversion in terms of interest also demonstrate similar aversion in terms of clicks, we employed two analytical methods.

First, using the Conditional GATES method (as described earlier), we examined whether

Figure 3.2: Group Average Treatment Effects

Figure 3.3: GATES on interest rate

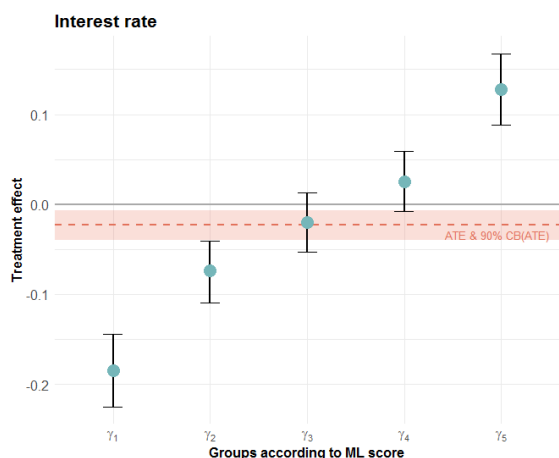
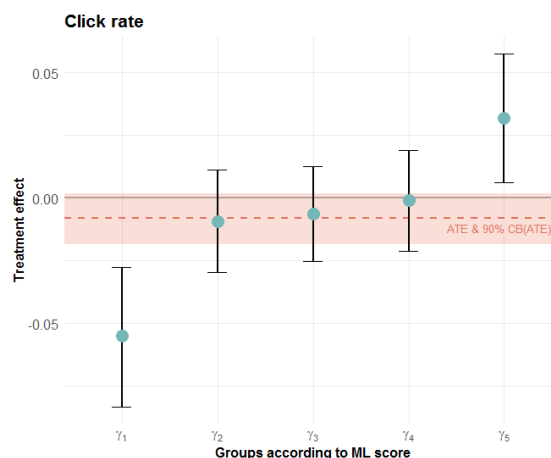


Figure 3.4: GATES on click rate



Note: This figure presents the Group Average Treatment Effects (GATES) for the interest rate (a) and click rate (b) outcomes. Individuals are divided into quintiles based on the intra-individual mean of the ML proxy, $\bar{S}(X_i)$. The vertical axis shows the treatment effect, and each point corresponds to the average treatment effect within each quintile, γ_k , with error bars representing 90% confidence intervals. The dashed line indicates the overall average treatment effect (ATE) for reference. The points and confidence intervals reported are the medians over 100 splits in half of the dataset. Light GBM is used as the ML proxy of the CATE.

Table 3.6: Group Average Treatment Effects

	Average Effect	Bottom 20% (Algorithm-Averse)	Top 20% (Algorithm-Friendly)	Top - Bottom
Interest (control mean: 0.18)				
Treatment effect	-0.023	-0.185	0.128	0.307
Confidence bands	[-0.039, -0.007]	[-0.225, -0.144]	[0.088, 0.167]	[0.252, 0.362]
Adjusted p-value	<0.01**	<0.001***	<0.001***	<0.001***
Click (control mean: 0.07)				
Treatment effect	-0.008	-0.055	0.032	0.087
Confidence bands	[-0.018, 0.002]	[-0.083, -0.028]	[0.006, 0.057]	[0.048, 0.124]
Adjusted p-value	0.216	<0.001***	<0.05*	<0.001***

Note: Medians over 100 splits in half. Light GBM is used as the ML proxy of the CATE. *, **, ***: significance at 10%, 5% and 1%.

the heterogeneity in treatment effects on the interest rate predicts the heterogeneity on the click rate. Table 3.7 presents the estimated treatment effects on the click rate for observations belonging to the algorithm-averse and algorithm-friendly groups based on the

interest rate. The results indicate that individuals in the bottom 20% (algorithm-averse group based on interest rate) exhibit a significant negative treatment effect on the click rate (-0.045), with a confidence interval of $[-0.071, -0.018]$ and an adjusted p-value less than 0.01. Conversely, individuals in the top 20% (algorithm-friendly group based on interest rate) show a significant positive treatment effect on the click rate (0.032), with a confidence interval of $[0.007, 0.056]$ and an adjusted p-value less than 0.05. The difference between the top and bottom groups is 0.075 , which is statistically significant (p-value < 0.001). These findings suggest that individuals who are algorithm-averse in terms of expressing interest also tend to be algorithm-averse in terms of clicking on job recommendations. Similarly, those who are algorithm-friendly in terms of interest are more likely to click on algorithm-labeled recommendations. This suggests a positive correlation between aversion in interest and aversion in clicks.

We also investigated the alignment between individuals' treatment effects on interest and clicks by analyzing the distribution of individuals across quintiles for both outcomes. For each data split, we trained our ML model on the auxiliary sample to estimate individual treatment effects on both the interest rate and the click rate. This provided us with two sets of predictions: S_{ij}^{int} and S_{ij}^{click} . In the main sample, we sorted individuals into quintiles based on their predicted treatment effects for each outcome. We then computed the distribution of individuals from each interest rate quintile across the click rate quintiles. Table C.6 shows the median of these distributions across all splits. The table shows that individuals classified as algorithm-averse based on their interest rate treatment effects are more likely to also be in the algorithm-averse quintile for click rate. Specifically, 28% of individuals in the first interest quintile are also in the first click quintile, which is higher than the expected 20% if there were no association. Similarly, individuals in the top interest quintile are more likely to be in the top click quintile (29%).

Table 3.7: Group Average Treatment Effects (quintiles for interest on clicks)

	Average Effect	Bottom 20 % (Algorithm-Averse)	Top 20% (Algorithm-Friendly)	Top - Bottom
Click (control mean: 0.07)				
Treatment effect	-0.008	-0.045	0.032	0.075
Confidence band	$[-0.018, 0.002]$	$[-0.071, -0.018]$	$[0.007, 0.056]$	$[0.039, 0.110]$
Adjusted p-value	0.216	$<0.01^{**}$	$<0.05^*$	$<0.001^{***}$

Note: Medians over 100 splits in half. Light GBM is used as the ML proxy of the CATE. *, **, ***: significance at 10%, 5% and 1%.

3.3.5 Characteristics of Algorithm-Friendly and Algorithm-Averse job seekers

Given the significant variation in how algorithm aversion affects job seekers, we aim to explore the characteristics and patterns associated with observations that fall into the algorithm-averse and algorithm-friendly categories. Our goal is not to make causal claims about specific traits that may directly influence this aversion. Instead, we focus on analyzing which characteristics are over-represented in observations classified into these two groups, providing a detailed overview of the patterns observed in the data.

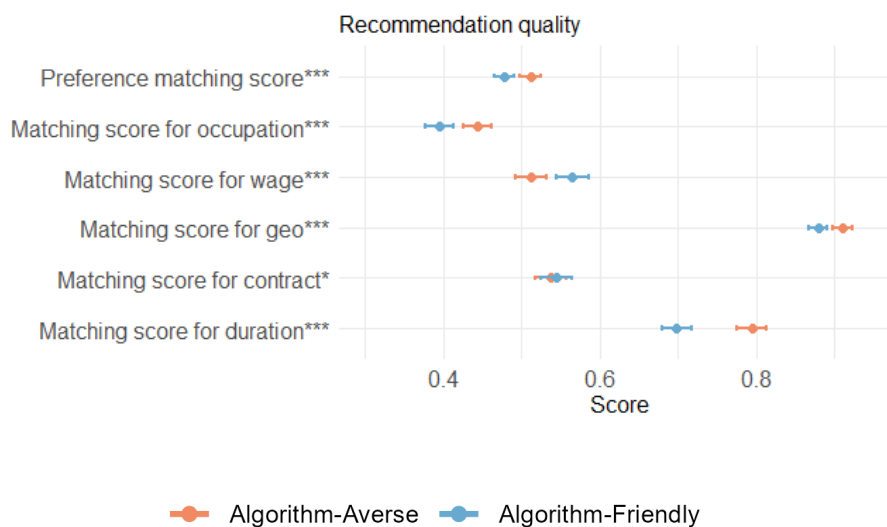
We present the results focusing solely on the click outcome, both to keep the analysis concise and, more importantly, because it is the most policy-relevant outcome. By concentrating on clicks, which represent actual engagement with job recommendations, we provide insights that are directly applicable to policy interventions aimed at improving job matching efficiency. The analysis on the interest outcome is provided in the appendix for completeness.

To provide a detailed overview, we split the results into subsections according to the categories of variables we employed to predict potential heterogeneity: variables at the jobseeker–recommendation pair level, describing the quality of the recommendations made, and variables defined at the jobseeker level, characteristics of their unemployment episodes, characteristics of the labor market in which they are searching. We created subcategories from all continuous variables defined at the jobseeker level to reveal clearer patterns and to analyze the distributions more finely than merely comparing means between the two groups.

It is important to emphasize that these findings represent correlations rather than causal relationships: each observed association might be proxying for another underlying factor. To address concerns about potential confounding variables, we assessed the strength of the associations between all pairs of covariates investigated below using Cramer’s V. The resulting matrix, presented in Figure C.6, documents relatively weak associations between the variables studied below, without any particularly strong patterns emerging.

Recommendations quality We begin by analyzing how the recommendation quality (i.e. how job recommendations align with predefined job seeker search criteria) varies between algorithm-averse and algorithm-friendly job seekers. Figure 3.5 reveals that the preference matching score is significantly higher for the algorithm-averse group, indicating that their job recommendations generally align more closely with a broader set of their specified preferences. In particular, the algorithm-averse group benefits from stronger matches in terms of job duration, geographic location, and occupation, suggesting that these individuals receive more tailored recommendations that meet their specific

Figure 3.5: Classification analysis on **Click rate**
Recommendation quality



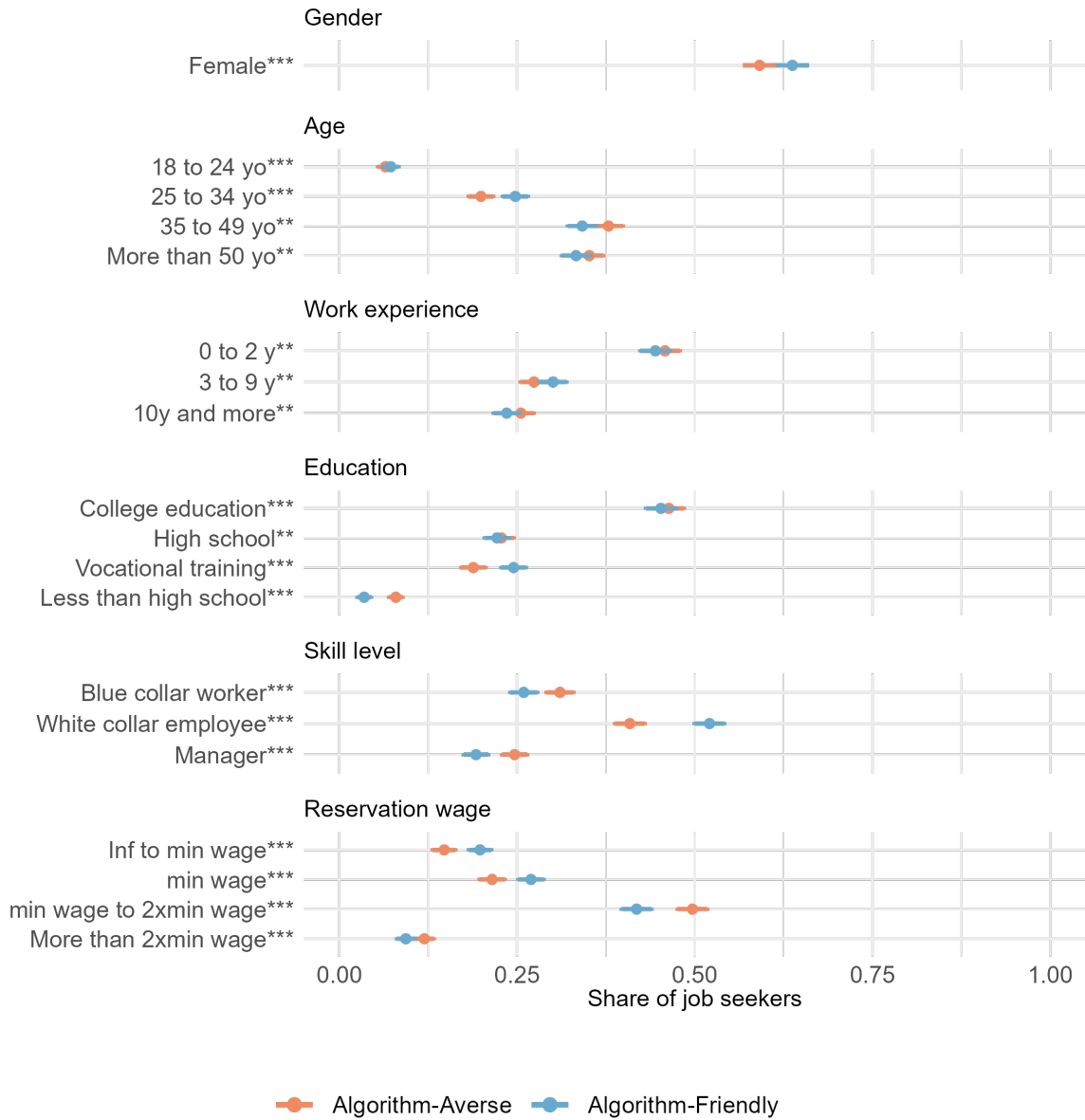
Note: The figure illustrates the differences in matching scores between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange) of job seekers. The horizontal axis represents a score ranging from 0 to 1. Higher values indicate a closer alignment between job seekers' preferences and the job recommendations they receive. Stars next to the variable names denote the significance level of the differences between the groups, with * for $p < 0.1$, ** for $p < 0.05$, and *** for $p < 0.01$. The dots indicate the median estimates derived from 100 random splits, while the lines represent the 90% confidence intervals. LightGBM serves as the machine learning proxy for estimating the Conditional Average Treatment Effect (CATE).

search criteria. One notable difference is observed in wage preferences: the algorithm-friendly group receives job recommendations more aligned with their wage expectations, contrasting with other matching criteria where the algorithm-averse group sees stronger alignment.

User characteristics The average user characteristics are displayed in Figure 3.8 and reveal notable demographic and socioeconomic patterns within the algorithm-averse and algorithm-friendly groups. First, female job seekers are significantly more common in the algorithm-friendly group, suggesting that women may be more open to algorithmic job search tools. This finding is consistent with some research suggesting women prefer algorithmic evaluators due to perceived fairness, especially in male-dominated settings [Pethig and Kroenung, 2023].

Age also plays a significant role in receptiveness to algorithmic assistance. The average ages are 43.6 for the algorithm-averse group and 41.9 for the algorithm-friendly

Figure 3.8: Classification analysis on **Click rate**
User characteristics

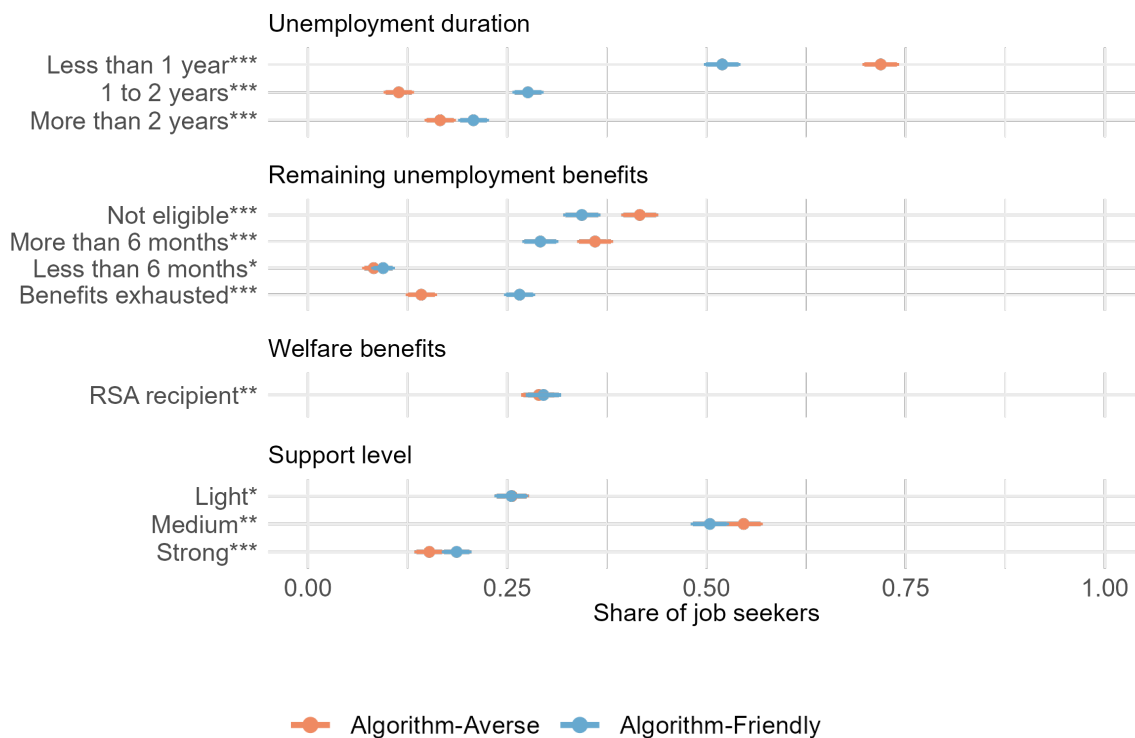


Note: This figure displays the distribution differences of job seekers' characteristics between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. LightGBM serves as the machine learning proxy for estimating the CATE.

group, showing a statistically significant difference ($p < 0.001$). More precisely, younger job seekers (aged 18–34) are more often found in the algorithm-friendly group, while job seekers over 35 are more often present in the algorithm-averse group. The age-related pattern that we find aligns with work experience: the average work experience in months is 87.6 for the algorithm-averse group and 84.8 for the algorithm-friendly group, also showing a significant difference ($p = 0.002$). This finding aligns with research indicating that older individuals are less trusting of algorithms and prefer human recommendations [Araujo et al., 2020, Thurman et al., 2019]. Older individuals may be more averse to algorithms due to their lack of familiarity with these tools, as previous research indicates that familiarity with algorithms can lead to greater acceptance [Fenneman et al., 2021].

Regarding education, there is a complex relationship between educational attainment and attitudes toward algorithmic recommendations. Prior research suggests that individuals with lower educational achievement levels and those less comfortable with numbers tend to appreciate algorithms less [Thurman et al., 2019, Logg et al., 2019]. In this study, job seekers with college degrees or with very low education levels (less than high school) are more prevalent in the algorithm-averse group. In contrast, individuals possessing vocational training are over-represented in the algorithm-friendly group. Similar patterns occur when considering job category: managers and blue-collar workers are more prevalent in the algorithm-averse group, while white-collar employees are more often represented in the algorithm-friendly group. This aligns with patterns in reservation wages: the average monthly reservation wages are significantly different, with the algorithm-averse group looking for positions offering an average of 2369.9 euros, while the algorithm-friendly group looks for positions offering an average of 2201.6 euros ($p < 0.001$). More precisely, lower wage earners (at or below the minimum wage) are more common in the algorithm-friendly group. Conversely, jobseekers with higher reservation wages (more than twice the minimum wage) are more frequently in the algorithm-averse group.

Figure 3.11: Classification analysis on **Click rate**
Unemployment spell characteristics



Note: This figure displays the distribution differences of job seekers' unemployment spell characteristics between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. LightGBM serves as the machine learning proxy for estimating the CATE.

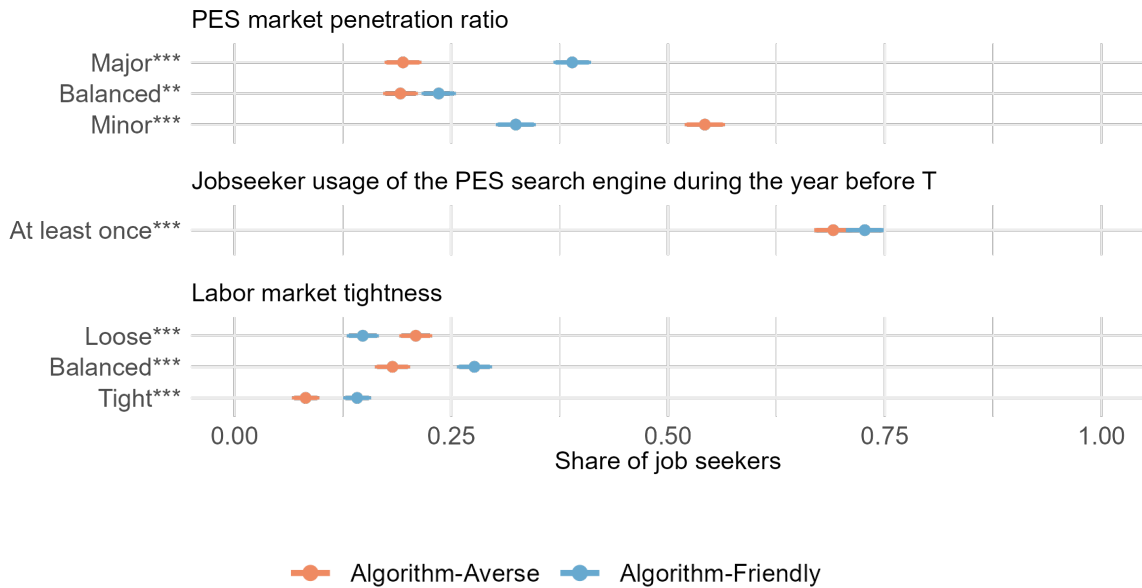
Unemployment-related characteristics Our analysis of unemployment-related characteristics reveals substantial variation between the algorithm-averse and algorithm-friendly groups. The average unemployment duration is notably longer in the algorithm-friendly group (16.2 months) compared to the algorithm-averse group (11.4 months) ($p < 0.001$). Figure 3.11 further reveals that job seekers with shorter unemployment durations (less than one year) are disproportionately represented in the algorithm-averse group, while job seekers with longer unemployment spells (over one year) are more common in the algorithm-friendly group. This pattern suggests that extended periods of unemployment

may lead individuals to prefer algorithmic recommendations over human ones. Similarly, job seekers nearing benefit exhaustion or with less than six months of remaining benefits are over-represented in the algorithm-friendly group. In contrast, job seekers with longer benefit coverage or who are ineligible for benefits tend to be in the algorithm-averse group. This indicates that financial urgency might influence job seekers to favor algorithmic tools over human assistance, perhaps due to a perception of efficiency or immediacy in algorithmic recommendations. Our findings resonate with [Marinescu and Skandalis \[2021\]](#), who highlight changes in job search behavior as unemployment benefits approach exhaustion. In their study, job seekers significantly increase their search efforts near benefit exhaustion, aiming to secure employment before benefits run out. This suggests that the financial urgency experienced by individuals close to exhausting their benefits might drive them to prefer the perceived efficiency of algorithmic recommendations. Additionally, the longer unemployment durations observed in the algorithm-friendly group align with the pattern identified by [Marinescu and Skandalis \[2021\]](#), where individuals adjust their job search strategies over extended periods. This suggests that as their job search intensifies, the need for immediate employment may outweigh the initial preference for traditional, human-based methods, potentially making job seekers more open to algorithmic assistance.

Regarding support from the PES, the level of caseworker assistance correlates with openness to algorithmic recommendations. Job seekers who receive intensive caseworker support are more prevalent in the algorithm-friendly group, whereas job seekers with medium support are more frequent in the algorithm-averse group. Job seekers with minimal support show no clear pattern. This may indicate that job seekers who have received intensive human assistance for an extended period feel discouraged by traditional tools they have used extensively, and thus view algorithmic recommendations as a more promising or novel approach.

Labor market conditions A key variable in our analysis is the market penetration ratio of the PES, which we define as the proportion of job postings published by recruiters on the PES platform relative to the total number of job postings in a given labor market, including postings on platforms not specifically dedicated to job seekers. This ratio serves as an indicator of how extensively recruiters rely on the PES to source candidates. Based on [Figure 3.14](#), compared to those in the algorithm-averse group, algorithm-friendly job seekers belong more often to markets where the PES is more frequently or equally used to advertise positions compared to other channels, and belong less often to markets where the PES is less influential. This pattern may be explained by several factors. Job seekers in PES-dominant markets may have had more exposure to the PES's digital tools and al-

Figure 3.14: Classification analysis on **Click rate**
Labor market conditions



Note: This figure displays the distribution differences of job seekers' labor market conditions between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. LightGBM serves as the machine learning proxy for estimating the CATE.

gorithms, fostering greater familiarity and trust in these systems compared to traditional, human-generated recommendations. In contrast, in markets where the PES has lower penetration, job seekers may be more accustomed to relying on other, non-algorithmic channels for job searches, thus making them more resistant to algorithmic recommendations. This aversion could stem from a perception that human caseworkers better understand the nuances of these less PES-reliant markets, leading to a preference for human-generated suggestions.

Labor market tightness (defined at the individual job seeker level) also appears to influence receptiveness to algorithmic recommendations. Job seekers in less competitive labor markets—where job vacancies are abundant relative to the number of jobseekers (classified as balanced or tight in the labor market tightness variable)—are over-represented

in the algorithm-friendly group. Conversely, jobseekers in highly competitive labor markets—where there are many jobseekers compared to available job ads—are more prevalent in the algorithm-averse group. This can be explained by the “scale effect” discussed in [Bana and Boudreau \[2023\]](#), where job seekers view AI-based recommendations as more scalable than human ones, and thus probably sent to other job seekers, fostering competition. As a result, they may perceive a higher level of competition when interacting with algorithmic tools, influencing their openness to algorithmic assistance. Additionally, this finding suggests that jobseekers facing less competition may feel more optimistic and open to exploring algorithmic assistance.

Conclusion

Our study contributes to the growing literature on the adoption and effects of algorithmic tools in labor markets, with a focus on the phenomenon of algorithm aversion. In a large-scale field experiment ($n = 30,000$) conducted with the French Public Employment Service, we randomize the perceived source of job recommendations—human experts, algorithmic generation, or no specific indication—to measure aversion to algorithms. The breadth of the surveyed population, coupled with the use of personalized, state-of-the-art recommendations directly relevant to job seekers’ job search, enhances the experiment’s ecological validity and policy relevance.

Our findings reveal significant average aversion among job seekers to algorithmic recommendations. On average, perceiving the recommendations as algorithmic, rather than generated by human experts, leads to a 10% decrease in declarations of interest (“thumbs up”) and a 15% reduction in clicks on recommended job ads. However, the study is under-powered to detect statistically significant effects on the number of job applications.

Leveraging rich administrative data collected by the PES, we investigate heterogeneity in responses to the perceived source of recommendations. We do not observe considerable heterogeneity in responses based on the quality of recommendations, as measured by their fit to job seekers’ explicit search criteria. To explore heterogeneity across high-dimensional demographic and socioeconomic characteristics, we employ the generic machine learning method of [Chernozhukov et al. \[2023\]](#) to flexibly and rigorously examine key features of the conditional average treatment effect function. Our analysis formally establishes the presence of heterogeneity in algorithm aversion. Despite the overall prevalence of aversion to algorithms, our results suggest the existence of algorithm-friendly subpopulations that tend to engage more with algorithmically labeled recommendations compared to those perceived as generated by human experts. We find that aversion to algorithmic recommendations is heavily influenced by demographic and socioeconomic

characteristics, including age, education, work experience, and reservation wages. Additionally, the context in which job seekers are searching—such as labor market tightness, unemployment duration, and the penetration of the PES platform—plays a significant role in shaping algorithm aversion.

While algorithmic recommendations and decision-making in labor markets offer potential efficiencies, our results suggest that algorithm aversion could pose a significant barrier to the effectiveness of these tools. This invites policymakers and Public Employment Services to consider behavioral responses when deploying such algorithmic systems. Notably, our findings indicate that improving the quality of recommendations alone may not suffice to mitigate algorithm aversion. Even when recommendations are well-matched to job seekers, resistance to algorithm-labeled suggestions persists. Therefore, attention should shift toward the framing of recommendations as a way to minimize the negative impact of algorithm labeling, as the presentation of recommendations—whether framed as algorithmic or human-curated—can significantly influence engagement. However, there is no universal framing solution that works for all job seekers. The heterogeneity in responses suggests that tailoring the framing based on individual characteristics may be a more effective approach. Alternatively, a neutral framing—where the source of the recommendation is not immediately disclosed, but job seekers have the option to learn more if desired—could potentially prove effective.

This study has several limitations. The survey's 20% participation rate introduced selection bias, as respondents differed from the broader job seeker population in terms of, for instance, familiarity with digital platforms. Additionally, while we varied the source framing (human, algorithmic, or neutral) to measure job seekers' responses, we cannot fully control how job seekers interpreted these labels. In all conditions, and especially in the neutral one, participants may have made their own assumptions about the source of the recommendations, potentially influencing their engagement and the measured levels of aversion. While a survey to collect participants' direct impressions of the perceived source could have clarified this, we gathered qualitative feedback on their satisfaction with the recommendations, perceived fit for job seekers' needs, and agreement on their chances of being hired for the suggested jobs, which will be used in future analyses. Finally, the study is under-powered to detect significant effects on later-stage outcomes, such as job applications and job placements.

References

- Aday. Intelligence artificielle dans les médias : l'effet chatgpt, 2023. URL <https://aday.fr/2023/06/14/intelligence-artificielle-dans-les-medias-effet-chatgpt/>. Accessed: 19/09/2024.
- Charu Aggarwal. Recommender Systems, volume 1. Springer, 2016.
- Steffen Altmann, Anita Glenny, Robert Mahlstedt, and Alexander Sebald. The direct and indirect effects of online job search advice. 2022.
- Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. In ai we trust? perceptions about automated decision-making by artificial intelligence. AI & society, 35:611–623, 2020.
- Sarah Bana and Kevin Boudreau. Behavioral responses to algorithmic matching: Experimental evidence from an online platform. 2023.
- Luc Behaghel, Sofia Dromundo, Marc Gurgand, Yagan Hazard, and Thomas Zuber. The potential of recommender systems for directing job search: A large-scale experiment. 2024.
- James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb):281–305, 2012.
- Guillaume Bied, Solal Nathan, Elia Perennes, Morgane Hoffmann, Philippe Caillou, Bruno Crépon, Christophe Gaillac, and Michèle Sebag. Toward job recommendation for all. In Thirty-Second International Joint Conference on Artificial Intelligence {IJCAI-23}, pages 5906–5914. International Joint Conferences on Artificial Intelligence Organization, 2023.
- Eric Bogert, Aaron Schechter, and Richard T Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. Scientific reports, 11(1):8028, 2021.
- Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.
- Gharad T Bryan, Dean Karlan, and Adam Osman. Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment. Technical report, National Bureau of Economic Research, 2021.

- Michèle Bélot, Philipp Kircher, and Paul Muller. Do the long-term unemployed benefit from automated occupational advice during online job search? 2022.
- Michèle Bélot, Philipp Kircher, and Paul Muller. Providing advice to jobseekers at low cost: An experimental study on online advice. The review of economic studies, 86(4): 1411–1447, 2019.
- Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. Journal of Marketing Research, 56(5):809–825, 2019.
- Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Fisherschlutz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india, 2023. URL <https://arxiv.org/abs/1712.04802>.
- Hyesun Choung, John S Seberger, and Prabu David. When ai is perceived to be fairer than a human: Understanding perceptions of algorithmic decisions in a job application context. International Journal of Human–Computer Interaction, pages 1–18, 2023.
- Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1):114, 2015.
- European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, April 2021. COM(2021) 206 final.
- Achiel Fenneman, Joern Sickmann, Thomas Pitz, and Alan G Sanfey. Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds. Plos one, 16(2):e0247084, 2021.
- Yoyo Tsung-Yu Hou and Malte F Jung. Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–25, 2021.
- Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. 2020.
- Guolin Ke, Qiwei Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qi Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems, pages 3146–3154, 2017.

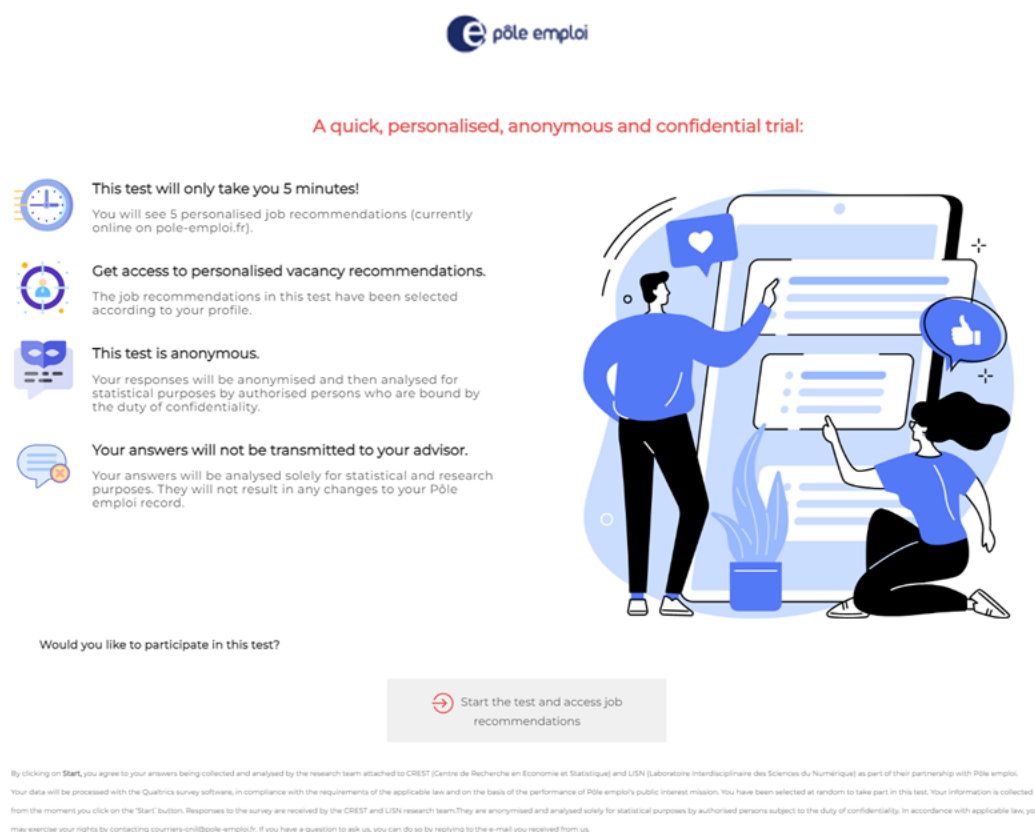
- Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J König, and Nina Grgić-Hlača. “look! it’s a computer program! it’s an algorithm! it’s ai!”: Does terminology affect human perceptions and evaluations of algorithmic decision-making systems? In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, pages 1–28, 2022.
- Sven Laumer, Fabian Gubler, Christian Maier, and Tim Weitzel. Job seekers’ acceptance of job recommender systems: Results of an empirical study. 2018.
- Thomas Le Barbanchon, Lena Hensvik, and Roland Rathelot. How can ai improve search and matching? evidence from 59 million personalized job recommendations. Technical report, Working Paper, 2023.
- Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1): 2053951718756684, 2018.
- Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90–103, 2019.
- Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4):629–650, 2019.
- Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175:121390, 2022.
- Ioana Marinescu and Daphné Skandalis. Unemployment insurance and job search behavior. The Quarterly Journal of Economics, 136(2):887–931, 2021.
- Yoosof Mashayekhi, Nan Li, Bo Kang, Jeffrey Lijffijt, and Tijn De Bie. A challenge-based survey of e-recruitment recommendation systems. ACM Computing Surveys, 56(10): 1–33, 2024.
- Jessica Ochmann, Leonard Michels, Sandra Zilker, Verena Tiefenbeck, and Sven Laumer. The influence of algorithm aversion and anthropomorphic agent design on the acceptance of ai-based job recommendations. In ICIS, 2020.
- Benjamin A Olken. Promises and perils of pre-analysis plans. Journal of Economic Perspectives, 29(3):61–80, 2015.

- Florian Pethig and Julia Kroenung. Biased humans, (un) biased algorithms? Journal of Business Ethics, 183(3):637–652, 2023.
- Andrew Prael and Lyn Van Swol. Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting, 36(6):691–702, 2017.
- Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. My friends, editors, algorithms, and I: Examining audience attitudes to news selection. Digital journalism, 7(4):447–469, 2019.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.
- Patrick van Esch, J Stewart Black, and Denni Arli. Job candidates’ reactions to ai-enabled job application processes. AI and Ethics, 1:119–130, 2021.
- Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. MIS quarterly, pages 425–478, 2003.
- Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In Proceedings of the 2020 CHI conference on human factors in computing systems, pages 1–14, 2020.
- Jenny S Wesche and Andreas Sonderegger. Repelled at first sight? expectations and intentions of job-seekers reading about ai selection in job advertisements. Computers in human behavior, 125:106931, 2021.
- World Bank. The use of advanced technology in job matching platforms: Recent examples from public agencies. 2023. URL <https://thedocs.worldbank.org/en/doc/ceb5c5792ad0d874e9b1c3cc71362f46-0460012023/original/Digital-Job-Matching-Platforms-S4YE-Draft-Note-for-Discussion.pdf>.
- Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. Journal of Behavioral Decision Making, 32(4):403–414, 2019.

Appendices

A The online survey

Figure A.1: Landing page



Note: At the bottom of the landing page, the disclaimer states: “By clicking on Start, you agree to your responses being collected and analyzed by the research team affiliated with CREST (Centre de Recherche en Economie et Statistique) and LISN (Laboratoire Interdisciplinaire des Sciences du Numérique) as part of their partnership with Pôle emploi. Your data will be processed using the Qualtrics survey software, in compliance with the applicable laws and based on Pôle emploi’s public interest mission. You have been randomly selected to participate in this test. Your information will be collected from the moment you click on the ‘Start’ button. The survey responses are received by the CREST and LISN research team, anonymized, and analyzed solely for statistical purposes by authorized individuals bound by confidentiality obligations. In accordance with applicable law, you may exercise your rights by contacting courriers-cnil@pole-emploi.fr. If you have any questions, you can respond to the email you received from us.”

Figure A.2: First page

Our recommendation algorithm has selected 5 job vacancies that might suit your profile.

Variation of the displayed source:
3 conditions among:
Algorithm / Human / Neutral
Condition « algorithm » is displayed here.

List of 5 recommendations:
with criteria:
- Occupation
- Company
- Town
- Work conditions
- Salary
- Required work experience
- Required diploma
- Required driving license

Your recommended vacancies:

Vineyard worker Job 1
DOMAINE DES GRANDS CHEMINS · St Jean De Muzols

Conditions: Temporary contract (3 months), Full time
Salary: 10.57 euros per hour
Work experience: Entry level
Diploma: Not specified
Driving license: Not specified

I am interested Not for me


Fruit picker Job 2
EARL MAISONNAS · St Jean De Muzols

Conditions: Temporary contract (3 months), Full time
Salary: 10.57 euros per hour
Work experience: At least 2 years
Diploma: Not specified
Driving license: Not specified

I am interested Not for me

Note: On the first page, the user sees the source of the recommendations at the top (only the displayed source varies, not the actual source). Below, 5 recommended vacancies are shown. For each vacancy, the user is asked to indicate whether they are interested or not.

Figure A.3: Bottom of the first page

 What do you think of these job recommendations?

Once you have answered these questions, click on the 'Finish test' button to save your response and to access the pages of these vacancies on pole-emploi.fr !

I am happy with these job recommendations.


Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
-------------------	----------	---------------------------	-------	----------------

Among all the vacancies currently online, these are among those for which I have the best chances of being recruited.

Pas du tout d'accord	Pas d'accord	Ni d'accord, ni pas d'accord	D'accord	Tout à fait d'accord
----------------------	--------------	------------------------------	----------	----------------------


These recommendations take into account the expectations and needs of jobseekers.

Pas du tout d'accord	Pas d'accord	Ni d'accord, ni pas d'accord	D'accord	Tout à fait d'accord
----------------------	--------------	------------------------------	----------	----------------------

 Thank you for answering. Save your answers by clicking on the 'Finish the test' button. If you wish, you can then consult these vacancies on pole-emploi.fr.

Note: At the bottom of the first page, the user can provide feedback on the recommendations received. Responding to these questions is optional; the user can skip them and proceed to the next page. At the bottom, there is a "Finish the survey" button. Clicking this button saves the user's responses (responses are not saved if the user leaves the page before clicking) and redirects them to the final page.

Figure A.4: Second page










You can consult these 5 vacancies on pole-emploi.fr and apply if you wish by clicking on the buttons below.

A new tab opens as soon as you click on the "See more details on pole-emploi.fr" button.

Job 1 / 5

Vineyard worker

Job vacancy criteria:

-  Company: DOMAINE DES GRANDS CHEMINS (2 employees)
-  Conditions : Temporary contract (3 months), Full time
-  Salary : 10.57 euros per hour
-  Work place : St Jean De Muzols (3.3 kilometers)
-  Work experience : Entry level
-  Diploma : Not specified
-  Driving license : Not specified

[See more details on pole-emploi.fr](#)

Same list of 5 recommendations:
with clickable buttons to see more details on the vacancy on the PES platform

Note: The second page, which is the final page, displays the 5 recommended vacancies again. This time, the user can click on each vacancy to be redirected to the PES platform for more details. The user's clicks are recorded.

B Summary statistics

Table B.1: Balance check among full sample

Sociodemographics, unemployment spell characteristics and labor market conditions

	Neutral	Human	Algorithm	<i>p</i>
Age: 18-24	0.17	0.17	0.17	0.89
Age: 25-34	0.32	0.31	0.32	0.31
Age: 35-49	0.31	0.32	0.31	0.03
Age: 50+	0.21	0.20	0.21	0.54
Blue collar	0.37	0.36	0.37	0.63
Manager	0.15	0.15	0.15	0.85
White collar	0.45	0.46	0.45	0.68
Work experience: 0-2 years	0.55	0.54	0.54	0.12
Work experience: 3-9 years	0.28	0.29	0.28	0.28
Work experience: 10+ years	0.15	0.15	0.15	0.40
Foreigner	0.15	0.16	0.16	0.02
Female	0.52	0.53	0.52	0.31
Highest diploma: college	0.36	0.36	0.36	0.54
Highest diploma: high school	0.25	0.26	0.25	0.21
Highest diploma: less than high school	0.09	0.09	0.09	0.45
Highest diploma: vocational	0.26	0.25	0.25	0.14
Married	0.40	0.41	0.40	0.73
No child	0.60	0.59	0.59	0.50
Level of assistance received from the PES: Light	0.23	0.23	0.24	0.31
Level of assistance received from the PES: Medium	0.53	0.52	0.52	0.68
Level of assistance received from the PES: Strong	0.21	0.21	0.20	0.35
Labor market tightness: Balanced	0.22	0.22	0.22	0.83
Labor market tightness: Loose	0.28	0.28	0.29	0.19
Labor market tightness: Tight	0.43	0.43	0.42	0.95
Duration of spell: 1-2 years	0.18	0.18	0.18	0.83
Duration of spell: Less than 1 year	0.64	0.65	0.65	0.23
Duration of spell: More than 2 years	0.18	0.18	0.18	0.30
Used the PES search engine in last 6 months	0.42	0.42	0.42	0.91
PES's market penetration: Balanced	0.23	0.23	0.22	0.07
PES's market penetration: Major	0.33	0.33	0.33	0.82
PES's market penetration: Minor	0.37	0.38	0.39	0.04
N	10 000	10 000	10 000	

Note: Columns (1) to (3) characterize job-seekers by their treatment assignment and report mean values (as a share of the sample unless stated otherwise); column *p* reports the *p*-value from the F-test for joint significance of treatment coefficients in the regressions of each covariate on treatment assignment.

Table B.2: Balance check among full sample

	Search criteria			
	Neutral	Human	Algorithm	<i>p</i>
High geographical mobility	0.18	0.18	0.18	0.21
Medium geographical mobility	0.65	0.65	0.65	0.67
Low geographical mobility	0.11	0.12	0.12	0.32
Monthly reservation wage: < min. wage	0.22	0.22	0.22	0.88
Monthly reservation wage: min. wage	0.27	0.27	0.26	0.69
Monthly reservation wage: min. wage - 2*min. wage	0.43	0.43	0.44	0.64
Monthly reservation wage: ≥ 2 *min. wage	0.07	0.07	0.07	0.53
Targeted job: Agriculture	0.03	0.03	0.03	0.74
Targeted job: Art & crafts	0.01	0.01	0.01	0.26
Targeted job: Banking & insurance	0.01	0.01	0.01	0.73
Targeted job: Business services	0.15	0.14	0.15	0.14
Targeted job: Comm, media & digital	0.02	0.02	0.02	0.13
Targeted job: Construction	0.06	0.07	0.06	0.09
Targeted job: Health	0.04	0.04	0.03	0.31
Targeted job: Industry	0.08	0.07	0.08	0.40
Targeted job: Maintenance	0.04	0.03	0.03	0.29
Targeted job: Performing arts	0.02	0.02	0.02	0.28
Targeted job: Personal services	0.16	0.16	0.17	0.19
Targeted job: Sales	0.13	0.15	0.14	0.04
Targeted job: Tourism & leisure	0.09	0.09	0.08	0.08
Targeted job: Transport	0.09	0.10	0.10	0.12
Looking for a full time job	0.65	0.65	0.64	0.41
Looking for a temporary contract	0.45	0.46	0.44	0.10
N	10 000	10 000	10 000	

Note: Columns (1) to (3) characterize job-seekers by their treatment assignment and report mean values (as a share of the sample unless stated otherwise); column *p* reports the p-value from the F-test for joint significance of treatment coefficients in the regressions of each covariate on treatment assignment.

Table B.3: Comparison of full sample and participants to the experiment

Sociodemographics, unemployment spell characteristics and labor market conditions

	Full sample	Respondents
Age: 18-24	0.17	0.08
Age: 25-34	0.31	0.22
Age: 35-49	0.31	0.35
Age: 50+	0.20	0.35
Blue collar	0.37	0.31
Manager	0.15	0.20
White collar	0.45	0.47
Work experience: 0-2 years	0.54	0.47
Work experience: 3-9 years	0.28	0.28
Work experience: 10+ years	0.15	0.24
Foreigner	0.16	0.12
Female	0.52	0.58
Highest diploma: college	0.36	0.42
Highest diploma: high school	0.25	0.22
Highest diploma: less than high school	0.09	0.07
Highest diploma: vocational	0.25	0.25
Married	0.41	0.49
No child	0.59	0.54
Labor market tightness: Balanced	0.22	0.23
Labor market tightness: Loose	0.29	0.28
Labor market tightness: Tight	0.43	0.43
Duration of spell: 1-2 years	0.18	0.18
Duration of spell: Less than 1 year	0.64	0.62
Duration of spell: More than 2 years	0.18	0.20
Receives intensive assistance from his/her caseworker	0.21	0.19
PES's market penetration: Balanced	0.22	0.22
PES's market penetration: Major	0.33	0.31
PES's market penetration: Minor	0.38	0.42
Used PES search engine in last 6 months	0.42	0.56
N	30000	6308

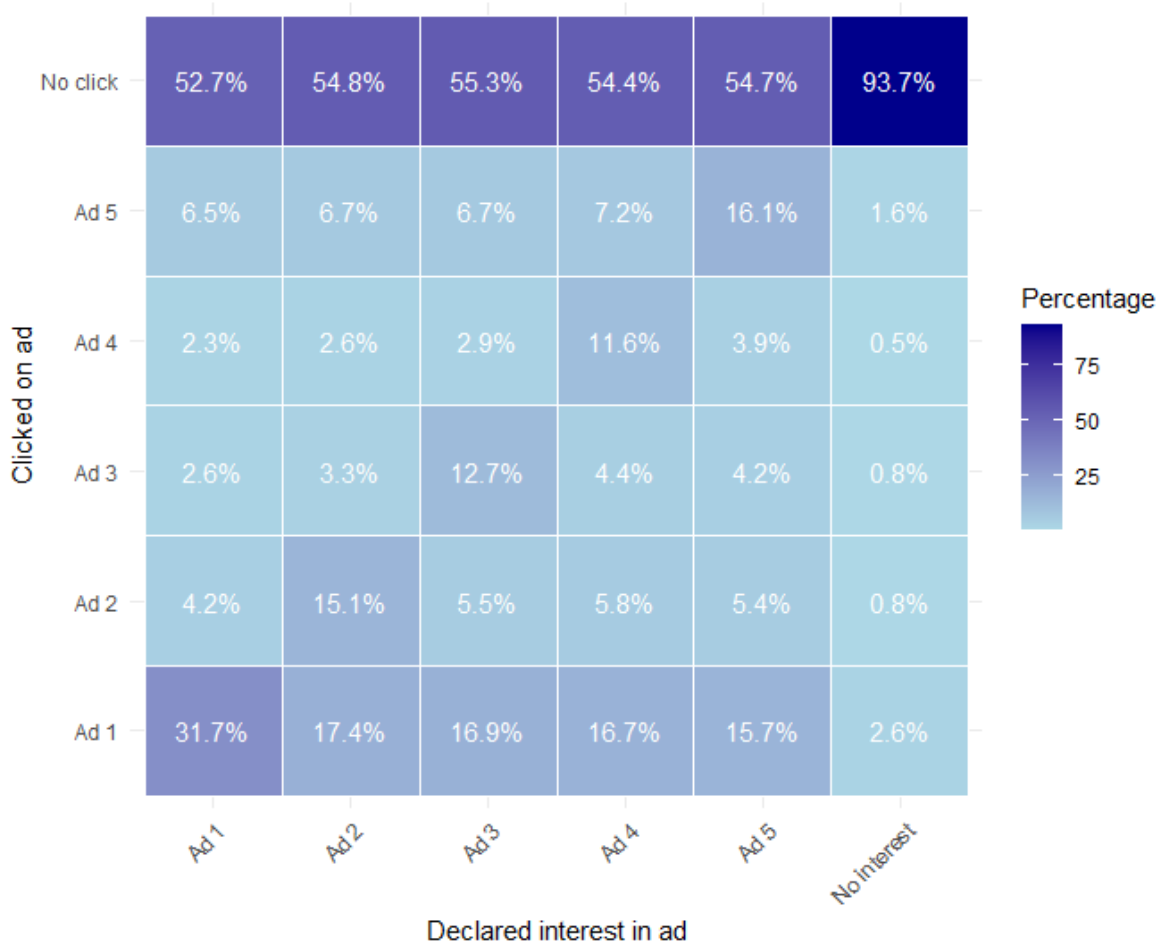
Note: Columns (1) and (2) report mean values (as a share of the sample unless stated otherwise) for the full sample population (column 1) and for the sample of participants who participated to the survey (column 2).

Table B.4: Comparison of full sample and participants to the experiment

Search criteria		
	Full sample	Respondents
High geographical mobility	0.18	0.18
Medium geographical mobility	0.65	0.65
Low geographical mobility	0.12	0.11
Monthly reservation wage: < min. wage	0.22	0.19
Monthly reservation wage: min. wage	0.26	0.25
Monthly reservation wage: min. wage - 2*min. wage	0.43	0.44
Monthly reservation wage: ≥ 2 *min. wage	0.07	0.11
Targeted job: Agriculture	0.03	0.02
Targeted job: Art & crafts	0.01	0.01
Targeted job: Banking & insurance	0.01	0.02
Targeted job: Business services	0.14	0.20
Targeted job: Comm, media & digital	0.02	0.02
Targeted job: Construction	0.06	0.05
Targeted job: Health	0.04	0.03
Targeted job: Industry	0.07	0.07
Targeted job: Maintenance	0.04	0.03
Targeted job: Performing arts	0.02	0.02
Targeted job: Personal services	0.16	0.18
Targeted job: Sales	0.14	0.13
Targeted job: Tourism & leisure	0.09	0.08
Targeted job: Transport	0.10	0.09
Looking for a full time job	0.65	0.62
Looking for a temporary contract	0.45	0.42
N	30000	6308

Note: Columns (1) and (2) report mean values (as a share of the sample unless stated otherwise) for the full sample population (column 1) and for the sample of participants who opened the survey (column 2).

Figure B.5: Association between declared interest and click behavior



Note: This heatmap illustrates the percentage of job seekers who clicked on ads relative to each ad they initially declared interest in. Each cell in the heatmap represents the proportion of job seekers, out of those who declared interest in a specific ad (indicated on the x-axis), who subsequently clicked on a given ad (indicated on the y-axis).

C Complements on results

C.1 Alternative specification for estimating the Average Treatment Effect

Taking into account the count nature of the outcomes The continuous outcomes we examine—representing the number of declared interests, clicks, and applications (each ranging from 0 to 5) for individual i —are count variables. Given the presence of overdispersion in the data, where the variance exceeds the mean, we employ a negative binomial model. The model assumes that the counts follow a conditional negative binomial distribution, with the logarithm of the conditional mean specified as a linear function of the regressors. Specifically, the conditional mean of the outcome, μ_i , is modeled as:

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot T_i^{\text{algo}} + \beta_2 \cdot T_i^{\text{neutral}} \quad (\text{C.1})$$

where T_i^{algo} and T_i^{neutral} are binary indicators for the treatment groups, capturing whether recommendations were labeled as originating from an algorithm or from a neutral source, respectively.

Table C.5: Impact of perceived source on the appreciation of recommendations, experts as reference, negative binomial model

Dependent variable	Interest (#)	Interest (Any)	Click (#)	Click (Any)	Application (#)	Application (Any)
(Intercept)	-0.119 (0.035)***	0.422 (0.011)***	-1.084 (0.048)***	0.235 (0.009)***	-3.251 (0.123)***	0.035 (0.004)***
Source: Algorithm	-0.112 (0.049)**	-0.025 (0.015)*	-0.150 (0.069)**	-0.027 (0.013)**	0.189 (0.167)	0.006 (0.006)
Source: Neutral	-0.051 (0.049)	-0.009 (0.015)	-0.102 (0.069)	-0.017 (0.013)	-0.007 (0.175)	-0.004 (0.006)
N. Obs.	6308	6308	6308	6308	6308	6308

Note: *, **, ***: significance at the 10%, 5%, and 1% levels.

Table C.5 presents the estimated effects of perceived source labeling on the number of declared interests, clicks, and applications using a negative binomial specification to account for the count nature and overdispersion of the outcomes. The results indicate that labeling recommendations as coming from an algorithm significantly reduces engagement for two key outcomes. Specifically, job seekers exposed to algorithm-labeled recommendations express approximately 10.6% fewer interests and 13.9% fewer clicks compared to those receiving expert-labeled recommendations, with both effects statistically significant at conventional levels. In contrast, the impact of neutral labeling is not significant for any of the outcomes. Interestingly, while the coefficient for algorithmic la-

being on the number of applications is positive, suggesting a potential increase of 20.8%, this effect is imprecisely estimated and not statistically significant. These findings suggest that while algorithmic labeling may reduce initial engagement (as measured by declared interests and clicks), it does not significantly alter the likelihood of job seekers submitting applications.

C.2 ML heterogeneous treatment effects

List of variables used to learn ML proxies

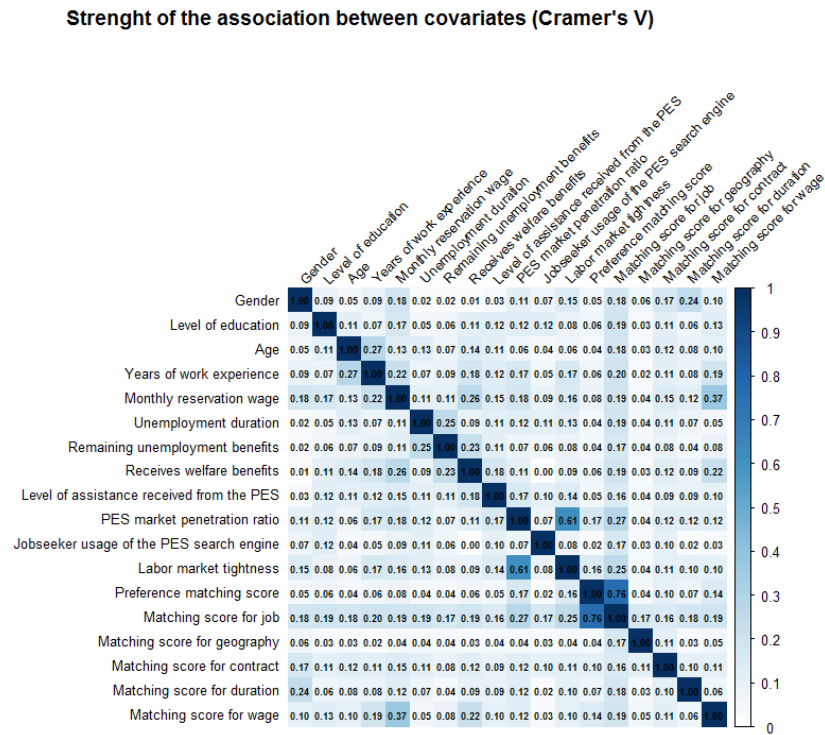
- Demographic information: age, gender, number of children, residence in sensitive areas, marital status and nationality.
- Education
- Work experience, measured in months of previous employment.
- Registration reason
- Support level
- Job category, describing the type of employment the individual seeks, such as managerial or blue-collar positions.
- Contract and work hours, distinguishing between permanent contracts and full-time employment.
- Unemployment duration, including the length of the unemployment spell and months of unemployment over the past year.
- Geographical mobility
- Occupation targeted
- PES market share and labor market tightness, reflecting the prevalence of job vacancies in the individual's area and the balance between job availability and competition.
- Reservation wage, indicating the individual's expected monthly wage.
- Benefits, including welfare benefits and the remaining eligibility for unemployment benefits.
- Matching scores, which represent how well an individual's profile matches various aspects of job recommendations, such as the job type, location, and wages.

- Utilization of the PES search tool, recording whether the individual has used the public employment service's search engine within the past year.

Table C.6: Flows between quintiles

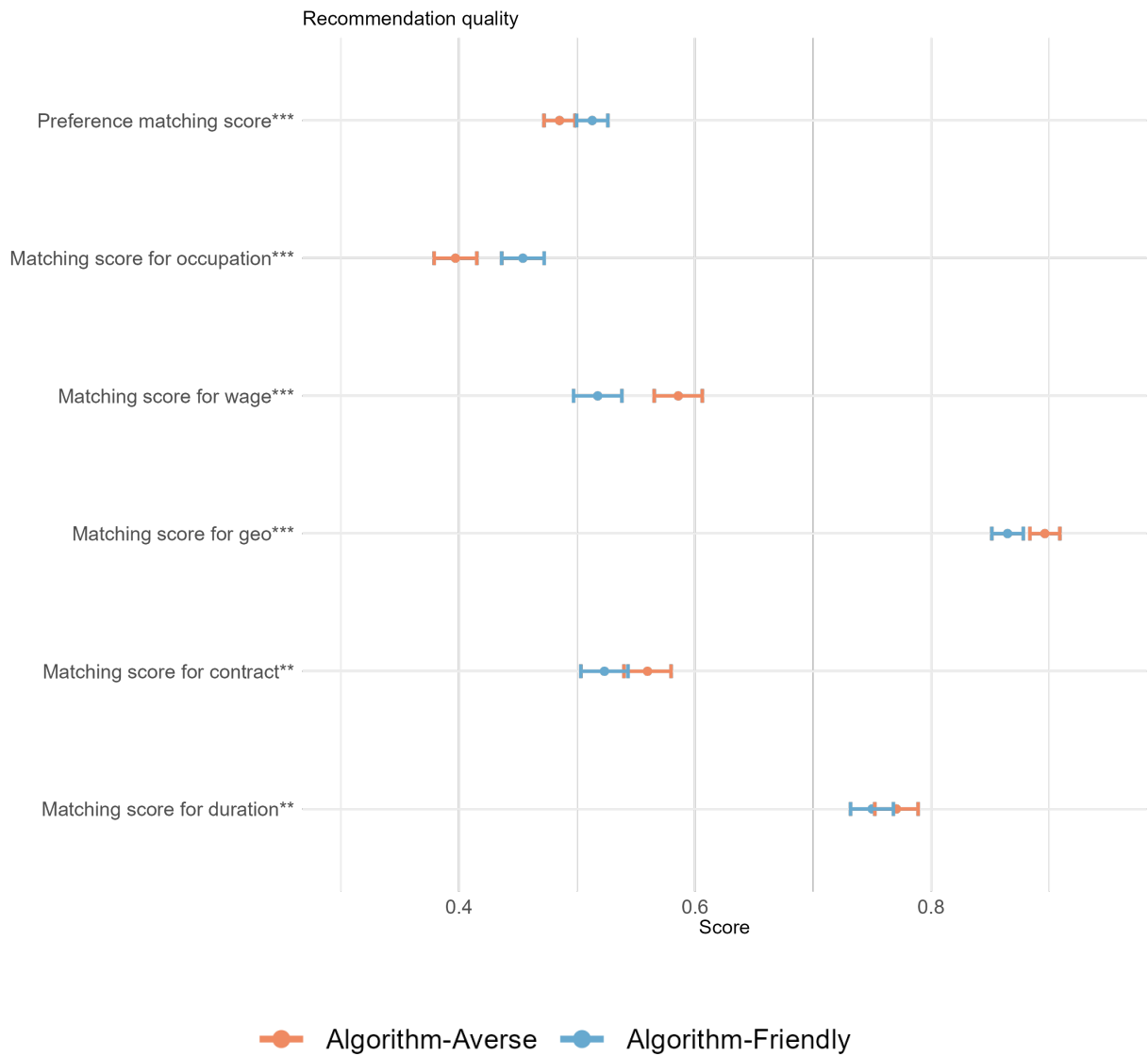
Group according to interest	Group according to click	Flow share (% of the interest quintiles)
G_1 (Algorithm-Averse)	G_1 (Algorithm-Averse)	0.28
	G_2	0.22
	G_3	0.19
	G_4	0.17
	G_5 (Algorithm-Friendly)	0.11
G_2	G_1 (Algorithm-Averse)	0.23
	G_2	0.22
	G_3	0.21
	G_4	0.19
	G_5 (Algorithm-Friendly)	0.16
G_3	G_1 (Algorithm-Averse)	0.19
	G_2	0.20
	G_3	0.21
	G_4	0.21
	G_5 (Algorithm-Friendly)	0.19
G_4	G_1 (Algorithm-Averse)	0.16
	G_2	0.19
	G_3	0.21
	G_4	0.22
	G_5 (Algorithm-Friendly)	0.22
G_5 (Algorithm-Friendly)	G_1 (Algorithm-Averse)	0.11
	G_2	0.16
	G_3	0.20
	G_4	0.23
	G_5 (Algorithm-Friendly)	0.29

Figure C.6: Strength of the association between the CLAN covariates



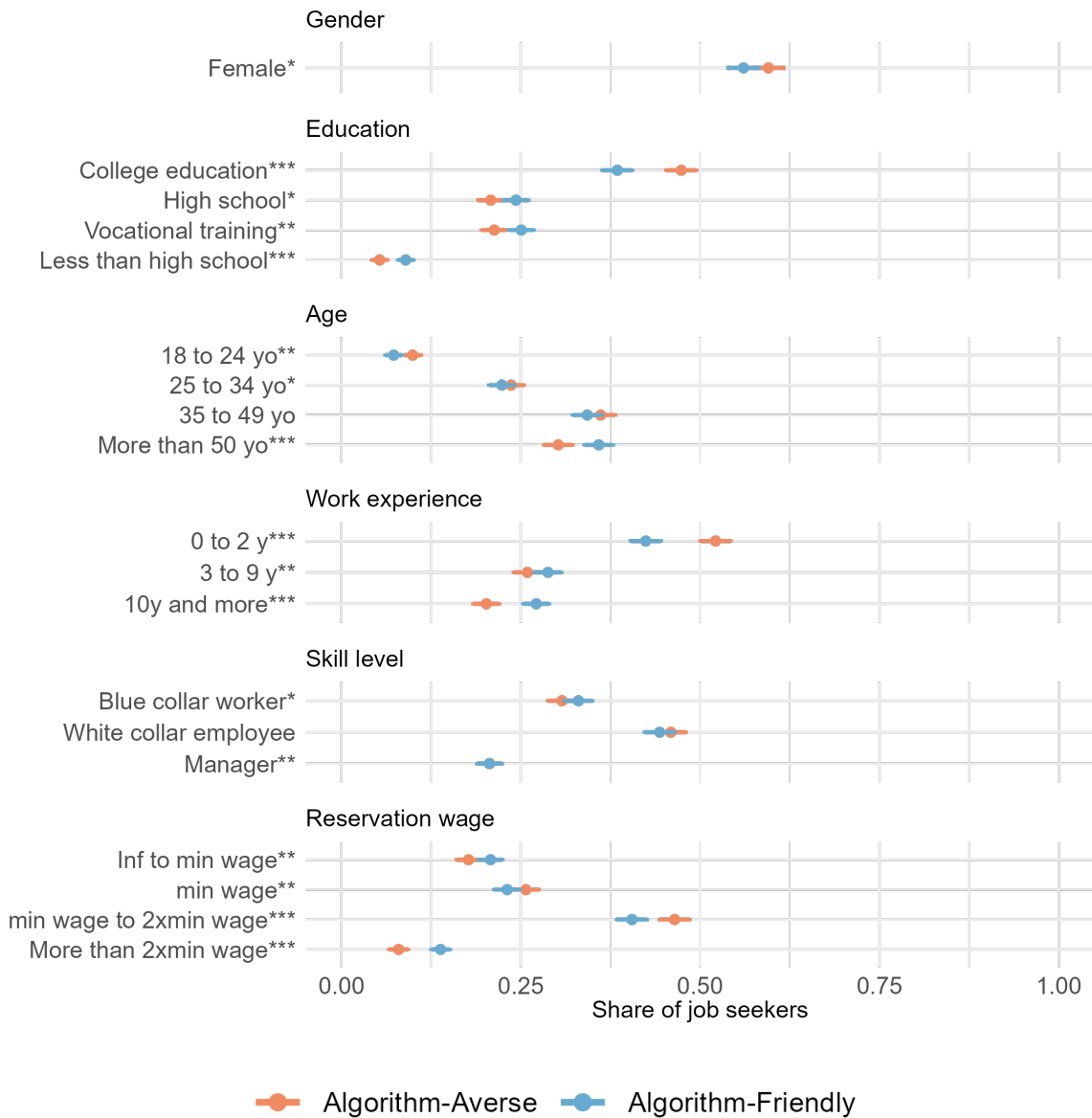
Note: This figure presents a heatmap of Cramér's V statistics, quantifying the strength of association between various categorical variables used in the analysis. Continuous variables (matching scores) have been transformed into categorical variables by taking quintiles. The values range from 0 (no association) to 1 (perfect association), where higher values indicate a stronger relationship between the corresponding pair of variables.

Figure C.7: Classification analysis on **Interest rate**
Recommendation quality



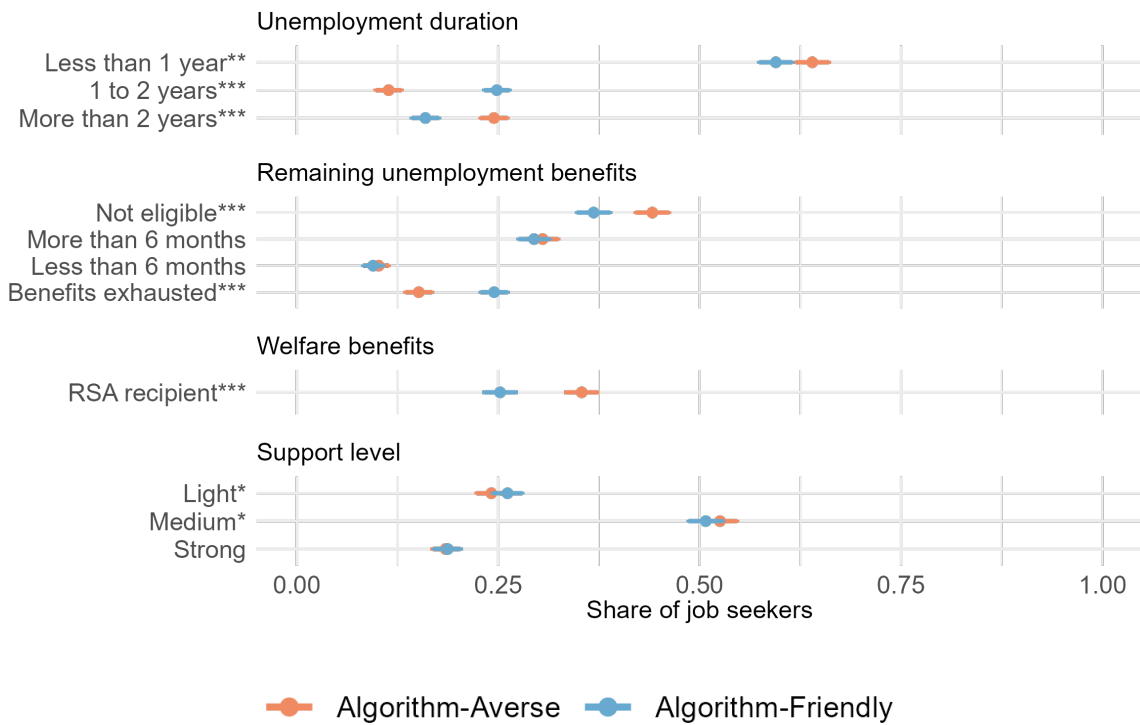
Note: The figure illustrates the differences in matching scores between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange) of job seekers. The horizontal axis represents a score ranging from 0 to 1. Higher values indicate a closer alignment between job seekers' preferences and the job recommendations they receive. Stars next to the variable names denote the significance level of the differences between the groups, with * for $p < 0.1$, ** for $p < 0.05$, and *** for $p < 0.01$. The dots indicate the median estimates derived from 100 random splits, while the lines represent the 90% confidence intervals. LightGBM serves as the machine learning proxy for estimating the Conditional Average Treatment Effect (CATE).

Figure C.8: Classification analysis on **Interest rate**
User characteristics



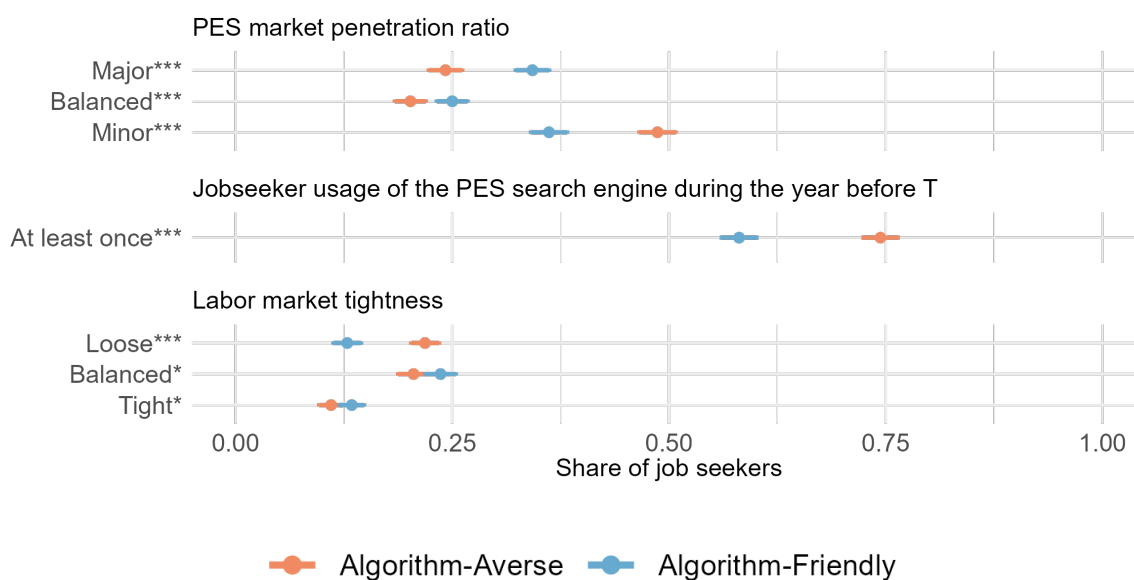
Note: This figure displays the distribution differences of job seekers' characteristics between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. Light-GBM serves as the machine learning proxy for estimating the CATE.

Figure C.9: Classification analysis on **Interest rate**
Unemployment spell characteristics



Note: This figure displays the distribution differences of job seekers' characteristics between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. Light-GBM serves as the machine learning proxy for estimating the CATE.

Figure C.10: Classification analysis on **Interest rate**
 Labor market conditions



Note: This figure displays the distribution differences of job seekers' characteristics between the algorithm-friendly group (in blue) and the algorithm-averse group (in orange). The horizontal axis measures the share of job seekers within each characteristic category, with higher values indicating a larger share. Stars are displayed alongside the variable names, indicating the significance level of the difference between the shares of the two groups (* for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$). The dots represent the median estimates derived from 100 random splits, while the lines indicate the 90% confidence intervals. Light-GBM serves as the machine learning proxy for estimating the CATE.

Titre : Conception et Évaluation des Systèmes de Recommandation sur le Marché du Travail

Mots clés : Recherche d'emploi sur Internet / Systèmes de recommandation / Apprentissage supervisé / Systèmes experts / Attitude devant ordinateur / Évaluation des politiques publiques

Résumé :

L'essor des systèmes de recommandation d'offres d'emploi constitue un enjeu important pour améliorer l'appariement entre demandeurs d'emploi et offres disponibles. Dans cette thèse, nous examinons, à travers trois chapitres complémentaires, les défis liés à leur conception, leur personnalisation et leur acceptation par les demandeurs d'emploi.

Dans un premier temps, nous comparons deux approches des systèmes de recommandation : l'une, basée sur l'utilité des demandeurs d'emploi, largement adoptée par les Services Publics de l'Emploi, évalue la compatibilité entre profils et critères des offres ; l'autre, fondée sur la probabilité d'embauche, repose sur des techniques d'apprentissage automatique pour prédire les appariements les plus susceptibles de conduire à une embauche. En intégrant ces deux dimensions dans un modèle économique validé empiriquement, nous démontrons que le système optimal, du point de vue des demandeurs d'emploi, combine utilité et probabilité d'embauche.

Nous examinons ensuite la personnalisation des systèmes de recommandation fondés sur l'utilité, en y intégrant les préférences individuelles des demandeurs d'emploi sur des attributs tels que le salaire,

la distance domicile-travail ou le type de contrat. À l'aide d'un essai contrôlé randomisé, nous montrons que cette personnalisation permet de mettre en lumière des offres que les demandeurs d'emploi auraient autrement manquées, entraînant une augmentation significative des clics et candidatures. Cependant, nous observons que le processus de recueil des préférences peut réduire les interactions avec les recommandations au niveau global, soulignant l'existence d'un arbitrage entre personnalisation et simplicité.

Enfin, nous analysons l'acceptation des recommandations algorithmiques par les demandeurs d'emploi. À travers un essai contrôlé randomisé, nous montrons qu'en moyenne, les demandeurs d'emploi éprouvent une préférence pour des recommandations présentées comme émanant d'une réflexion humaine, par rapport à celles présentées comme algorithmiques. Nous distinguons ensuite deux segments de demandeurs d'emploi : les "favorables aux algorithmes" et les "averses aux algorithmes", différenciés par leurs caractéristiques socio-économiques. Ces résultats mettent en avant l'importance d'adapter la présentation des recommandations pour encourager leur adoption.

Title : Designing and Evaluating Labor Market Recommender Systems

Keywords : Online job search / Recommender systems / Machine learning / Expert systems / Human-computer interaction / Public policy evaluation

Abstract :

The development of job recommendation systems is an important issue to improve the matching between job seekers and available jobs. In this thesis, we examine the challenges of design, personalization, and job seeker acceptance of these systems in three complementary chapters.

First, we compare two approaches to recommender systems: one based on jobseeker utility, widely used by public employment services, evaluates the compatibility between profiles and offer criteria; the other, based on hiring probability, relies on machine learning techniques to predict the matches most likely to lead to hiring. By integrating these two dimensions into an empirically validated job search model, we show that the optimal system from a jobseeker's perspective combines both utility and hiring probability.

We then explore the personalization of utility-based systems by incorporating job seekers' individual pre-

ferences for attributes such as salary, commuting distance, or contract type. Using a randomized controlled trial, we show that this personalization brings to light otherwise ignored vacancies, leading to a significant increase in clicks and applications. However, we find that the process of collecting preferences can reduce overall interactions, highlighting the need for a trade-off between personalization and simplicity.

Finally, we analyze job seekers' acceptance of algorithmic systems. The results of a randomized controlled trial show an average preference for recommendations perceived as coming from human reflection over those identified as algorithmic. We distinguish two segments of users: the "algorithm-friendly" and the "algorithm-averse", differentiated by their socio-economic characteristics. These observations highlight the importance of adapting the presentation of recommendations to encourage their adoption.