UAB
Universitat Autònoma de Barcelona

Active labour market policies: effects and predictions for policy design

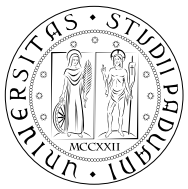
Álvaro Fernández Junquera

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons: https://creativecommons.org/licenses/?lang=ca

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons: https://creativecommons.org/licenses/?lang=es

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license: https://creativecommons.org/licenses/?lang=en

UAB
Universitat Autònoma
de Barcelona



Università
degli Studi
di Padova

Active labour market policies: effects and predictions for policy design

Álvaro Fernández Junquera

Doctoral thesis

Supervisors: Oscar Molina Romo and Enrico Rettore

Tutor: Oscar Molina Romo

Ph.D. program in Sociology
Universitat Autònoma de Barcelona

Ph.D. program in Economics and Management
Università degli Studi di Padova

Academic year 2023/2024

Department of Sociology (UAB)
Department of Economics and Management (UNIPD)

To the millions who suffer through want of work

To those who push to solve the unemployment problem

Acknowledgments

Doctoral studies are a long journey of four years in which the support of many people becomes crucial. I like the Ronald Fisher's concept of "research workers" to refer scholars. During these four years, we begin to learn how to do our work. Fortunately, the learning process is easier and more enjoyable when it is shared.

Before coming to Barcelona, I completed my bachelor's and master's degrees in the beautiful city of Granada. During my undergraduate studies, Prof. Francisco González encouraged me to pursue a research career and to explore economics, showing me the value of interdisciplinarity. During my graduate studies, Prof. Mª Nieves Pérez trusted me and opened the doors to her master's program, even though I was a political science student. I am indebted to both of them.

Over the four years in Barcelona, I would like to highlight four persons who have been with me in the day-to-day activities of the research group: Alejandro Godino, Daniel Barrientos, Alejandro González, and Sander Junte. Thanks to all of them for making this journey more pleasant and more manageable. Another important person has been Josep Márquez, from the public employment service of Catalonia. His passion for policy evaluation has built a bridge from academia to public policy, pushing public administrations towards more evidence-based policy design. Thank you for your confidence in me.

In 2021, I had the opportunity to visit the University of Buenos Aires. I would like to thank all the members of IIEP for their warm welcome, but especially Martín Llada for his grateful support during my stay in Argentina. Other colleagues closer to politics made me believe that a more just and popular world is possible: Tomi, Félix, Diego... Thank you all!

In 2022 and 2023, I spent part of my doctoral studies at the University of Padua. There, I met wonderful people and I learnt how close Italy is to Spain. I would like to thank the doctoral students of Via del Santo for sharing their time with me. Colleagues from the public employment service of Veneto have made my work much easier and been very generous, so thank you very much as well. I also remember with special affection my friends from Medici con l'Africa – CUAMM. Stefano, Lisa, Giorgio, Fede, Silvia, Xabi, Coro... It was a real pleasure to live for seven months in a vibrant community of amazing people. *Grazie mille!*

In 2024, I visited the Ludwig Maximilian University of Munich. Prof. Christoph Kern played a key role during the last year of my doctoral journey and always trusted me. I am very grateful for his collaboration. During my stay in Germany, I felt fully integrated into the SODA group. Thanks to all of you and especially to my friend Clara Strasser for the coffees we shared.

Throughout the different stages of the doctoral thesis, my two supervisors (Prof. Oscar Molina and Prof. Enrico Rettore) have supported my research with their suggestions and encouragement. I have learnt a lot about labour markets and econometrics from both of them. Thank you very much.

Beyond academia, my family and friends have always backed my research through daily cares. I would like to thank my parents and my siblings for their love and support. Chiclana and Barcelona have also concentrated many hours of friendship. This thesis is also yours for the understanding and reception of the ups and downs of the journey. The four years have been more fun with you, so thank you!

Special thanks go to the four mathematicians who shared their passion for maths with me through numerous discussions at home. Thanks to Ángel, Nil, Pablo, and Fran. Ángel has also been a pillar during the four years of doctoral studies because of his help, listening, and laughs. *¡Gracias!*

Lastly, I would like to acknowledge the funding of the Ministry of Universities of the Government of Spain for a predoctoral position through the Plan de Formación del Profesorado Universitario (FPU) with the grant FPU19/05291.

Abstract

This doctoral dissertation aims to provide evidence for the design of active labour market policies (ALMPs). The research is structured into two main objectives: evaluating the effectiveness of specific design features of ALMPs on employment outcomes and assessing jobseeker profiling models to optimize the selection of target groups.

To address the first objective, we developed two impact evaluations. In Chapter II, we investigated the role of timing, intensity, and occupational specificity in an adult training program in Catalonia (Spain). We employed a conditional difference-in-differences strategy and found that the overall employment effect of the program was positive. However, we did not find clear evidence of a timing effect, as we obtained minor differences across cohorts. We detected a remarkable non-linear effect of the intensity of work-based learning (internships) on the employment probability. Nonetheless, the impacts of the intensity of classroom-based learning and occupational specificity were substantially smaller.

In Chapter IV, we analysed the role of job search assistance and intermediation in a program implemented through vouchers and performance-based pay in Veneto (Italy). The policy assigned individuals to three treatment groups with different treatment endowments according to the score of a profiling model. Using a regression discontinuity strategy, we compared the effects of different treatment levels and did not find an impact on the employment outcomes after two years. Regarding the assistance component, one key finding was that the vouchers were underutilized, leading to minimal differences in treatment uptake. Concerning the intermediation component, the differences between the payments of each treatment group seemed overall insufficient to yield lasting effects.

To achieve the second objective, we studied the predictive performance of two types of jobseeker profiling models. The rule-based model currently deployed in the public employment service of Catalonia was found to have poor predictive performance, both in its ability to discriminate jobseekers that will be long-term unemployed and in calibrating the employment predictions. In contrast, a newly developed statistical model demonstrated superior performance, leveraging

administrative data and machine learning techniques to achieve excellent calibration. This model could potentially better align ALMP resources with individual needs.

Our research has tried to contribute both to the academic and the policy spheres. With respect to the academic literature, we have paid greater attention to the specific tools and target groups included in program designs. This approach allows us to go beyond the simple classification of types of ALMPs, digging into the components of the intervention that might attain a behavioural or cognitive change. Moreover, we have advanced contemporary research on different topics like the occupational specificity or the voucher implementation by studying them in the framework of ALMPs. With respect to policymaking, we have brought evidence on highly discussed issues in Southern Europe like the profiling of jobseekers or the use of internships in adult training programs. Both elements have been recently regulated at the national level in Spain and Italy, so our analyses might serve to continue the policy design at the rest of government levels.

Contents

Chapter I. Introduction 10
1. Concepts 12
1.1. Public policies 12
1.2. Labour market policies 14
1.2.1. Tools of ALMP 19
1.2.2. Target groups of ALMP 28
2. Justification 30
2.1. Spain and Italy 30
2.2. International cotutelle 39
3. Objectives and hypotheses 40
3.1. Objectives 40
3.2. Hypotheses 41
4. Methodology 42
4.1. Common to all chapters 43
4.2. Specific to each chapter 45
Chapter II. Employment effects of adult training in Catalonia: heterogeneity by design 47
1. Introduction 47
2. Adult training 50
2.1. Literature review 50
2.2. Design of the program 53
3. Methods 55
3.1. Data 55
3.2. Analysis 57
3.2.1. Target parameters 57
3.2.2. Identification 58
3.2.3. Estimation and inference 63
4. Results 65
4.1. Time-based heterogeneity 65
4.2. Tool-based heterogeneity 73
4.3. Robustness analyses 82
5. Conclusions 88
Appendix 92
Chapter III. Evaluating two types of profiling models for the public employment service of Catalonia 108
1. Introduction 108

2. The Catalan profiling model in context 110
2.1. Jobseeker profiling 110
2.2. The Catalan profiling model 114
3. Methods 116
3.1. Data 116
3.2. Analysis 119
3.2.1. Model development 119
3.2.2. Model validation 121
3.2.3. Model selection and interpretation 125
4. Results 127
4.1. Performance comparison 127
4.2. Model similarity and interpretation 133
5. Conclusion 137
Appendix 140
Chapter IV. Employment effects of job search interventions in Veneto assistance and intermediation 159
1. Introduction 159
2. Job search services 161
2.1. Literature review 161
2.2. Design of the program 164
3. Methods 166
3.1. Data 166
3.2. Analysis 167
3.2.1. Target parameters 167
3.2.2. Identification 168
3.2.3. Estimation and inference 173
4. Results 173
4.1. Employment effects 173
4.2. Mechanisms 181
4.3. Plausibility and robustness 183
5. Conclusion 192
Appendix 195
Chapter V. Conclusions 215
References 219

Chapter I.

Introduction

Labour markets have been evolving in the last years due to international trends like technological progress, globalisation, and demographic change (OECD, 2019a). Firstly, technical changes in areas like automation or information and communication technologies have been suggested both as labour substitutes (e.g. industrial robots) and as labour complements (e.g. computers or phones). Nedelkoska and Quintini (2018) estimated that only 14 % of current jobs suffer the risk of full automation. However, the fear of job destruction has augmented with the development of artificial intelligence (OECD, 2023a). Secondly, the increase of international trade since 1975 has been clear in the advanced economies, but also in emergent economies like China or Eastern Europe (Donoso et al., 2015; Dauth et al., 2017). Nonetheless, Autor et al. (2016) show that the analyses on the benefits of globalisation for the consumers must be accompanied by the costs of slow labour adjustments in local labour markets. Thirdly, the population of developed countries like Spain or Italy is remarkably ageing while developing countries need to integrate many young people into labour markets. In the OECD, this may reinforce the higher demand of services (like cares) and the lower demand of durable goods (like cars), so changes in the sectorial structure are expected. Such adjustments might be assisted by more migration, but we must consider that foreign workers face larger labour market risks (OECD, 2019a).

These international trends might affect more severely countries whose national labour markets already presented important deficiencies. In this dissertation, we focus on Spain and Italy, two countries with the highest rates of people at risk of poverty or social exclusion in the EU (OECD, 2023c). In both countries, one out of four persons is at this situation. Moreover, the low decrease of these rates in most OECD countries suggest that this indicator might hidden structural problems (Eurostat, 2024f). However, social exclusion is a multidimensional concept and not all countries suffer from high unemployment. Spain and Italy do have a problem by presenting the largest rates of long-term unemployment *and* of in-work poverty in the EU (Eurostat, 2024c; Hick & Marx, 2023).

To address the risks and potentials of these international trends, public policies appear as useful instruments. Hemerijck (2017) summarizes with three words the three functions that social policy may play in contemporary societies: stocks, flows, and buffers. First, it may elevate the level of the *stock* of productivity (in market assets) or capabilities (in non-market assets) with which individuals face different risks along their life-course. For instance, adult training for employed workers updates them to use new technological advancements. Second, it may ease the *flow* from different labour market or life-course states. For example, policies like job search assistance speed up labour reallocation in import-exposed industries. Third, it may act as a *buffer* against the lack of sufficient incomes. With low fertility rates, family benefits may diminish the economic risks of having children and then the population ageing might stop. Importantly, policies like unemployment benefits may work on the three different areas in a complementary way: buffering the drop of income, facilitating a smooth flow from a well-matched to another well-matched job, and retaining an alive stock of human capital.

This doctoral thesis evaluates three interventions related to labour markets with the aim of improving the designs of active labour market policies. It is made up of three research chapters based on working papers authored or coauthored by us that were developed as part of the doctoral studies. The first policy is an adult training program implemented in the region of Catalonia (Spain) from 2016 to 2018. These calls had the largest budget in the training area of the Catalan public employment service (SOC, 2024). The second policy is the jobseeker profiling model deployed in the public employment service of Catalonia from 2015 to 2020. This policy is not aimed at changing the behaviour of jobseekers, but at predicting it with the objective of selecting the target groups of active labour market policies. The third policy is a program of job search services executed in the region of Veneto (Italy) from 2017 to 2022. This has been the regional instrument available for workers older than 30 years to fund the participation in the active labour market policies offered in the quasi-market of private providers.

The chapter continues in Section 1 with the definition of key concepts used during the doctoral thesis. Then, we offer a justification in Section 2 of the cases of study and the pertinence of an international cotutelle. Section 3 mentions the objectives and hypotheses of the investigation. Lastly, Section 4 describes the methods used across the whole thesis and those specific of each chapter.

1. Concepts

1.1. Public policies

Public policies have been briefly defined as "whatever governments choose to do or not to do" (Dye, 2012, p. 3). Maintaining the essence, Howlett (2019) relying on Lasswell (1951) defined public policies as "government decisions composed of policy goals and means at different levels of abstraction" (p. 4)[1]. The first element of this definition, the policy goals, regards the objectives the policymakers have in mind when carrying out some action or inaction. The second element specifies the tools or instruments used to reach such goals, which encompasses both non-interventions and interventions. The third element is especially important in decentralized countries like Spain or Italy, since it highlights the multi-level governance of a policy. The different levels of abstraction of goals and means are the product of different levels of government, being in a hierarchical way or recommended in a non-compulsory way[2].

The development of public policies is usually described through different tasks depicted in a policy cycle (Subirats et al., 2008). Inside this process, following Howlett & Mukherjee (2018), the design of a policy may be understood as one stage of the cycle (as a *verb*) or as the product of such stage (as a *noun*). In other words, the concept of policy design is used both to describe the content of a policy as reflected in official texts and the task of formulating it (Siddiki & Curley, 2022). Under this model, the stage of design is followed by the implementation of the policy and its impact evaluation. The key point is that policies usually survive more than one fiscal exercise, so the impact evaluation is then used for policy design. To make this possible, the evaluation should give some insights on the specific elements to include in a design.

---

[1] Two conceptual clarifications are pertinent. First, we will sometimes use "policy" to mean "public policy" for brevity. Notice that there are also private policies, but we focus on the public ones. Second, considering that all policies may be analysed at different levels of abstraction, we use policy, program, and intervention as synonyms.

[2] Classic studies in political science usually distinguish policy design from policy implementation (Subirats et al., 2008). The hierarchical approach of our definition involves that a decision will be considered part of the design or implementation *depending on* the agent regarded as central. For instance, an order issued by the manager of the public employment service is part of the design for such organism and part of the implementation for the parliament. The non-compulsory approach regards the distinction between hard and soft law that has made its way into many public policies, including the labour market policies (Pérez del Prado, 2021).

If we focus on the design as content, the seminal work of Schneider & Ingram (1997) tried to detect a set of elements common to many policies. It is possible to integrate the core of their list of elements into the goals and means architecture of our initial definition. Regarding the goals, it is necessary to specify what is the desired situation in a certain space and time with mention to the population of interest, i.e. the target groups (Sabatier & Mazmanian, 1980). Concerning the means, policymakers should pay attention to the type of tools to achieve the goals and to their specific features. There are three types of tools or instruments applied in many policy fields to influence the behaviour of the targeted individuals: goods/services provision, money transfer, and regulation (Howlett, 2018)[3]. We can even reduce this typology to two classes: financial tools (goods/services provision or money transfer) and nonfinancial tools (regulation). However, designing a policy is more intricate than selecting one of these three options. Some of the features of the design are transversal across different policy fields, like the transfer channel of money (e.g. voucher vs. cash) or the sector that manages a public service provision (e.g. with public employees or via outsourcing). Other features are more dependent on the policy field, like the occupations targeted by a training course. Therefore, it is not possible to make a complete list of design features, but some of them might be the legal form of the design documents, the duration, the timing, the legal form of the providers, the level of services or the amount of money offered (Howlett & Goetz, 2014; Howlett & Mukherjee, 2018).

There are two extra elements emphasized by Schneider & Ingram (1997) that should be present in a policy design: assumptions and rationales. Firstly, a good design includes a set of statements that acts as glue between means and goals: the *assumptions* on the causal model. This is usually depicted with a figure called theory of change, logic model, or causal graph. Visualizing the causal graphs of many policies allow us to cluster them by fields of public policy like health, education, or energy. Secondly, the *rationales* regard the justifications of existence of certain goals or means in an isolated way. They address questions like "why should we try to decrease part-time employment?" (on goals) or "why should we pay an unemployment insurance?" (on tools). Taking all these elements into consideration, we have a policy design.

[3] Howlett (2018) originally talks about a fourth group: information-based instruments. However, information provision may be considered a service and be integrated into the first group.

Following the cycle of the policy, after the design and implementation of a program, such policy design may be judged according to the social welfare it generates. On the means side of the design, the bad news is that the strength of the connection between means and goals is not obvious, so it must be estimated using evaluation techniques. In short, assumptions on a causal model are assumptions. Furthermore, public policies may even have unintended consequences on outcomes they did not target (Linares Martínez, 2018). On the goals side of the design, although the causal model captured existing effects, the groups targeted may not be the optimal ones for achieving the maximum social welfare.

1.2. Labour market policies

Our thesis focuses on active labour market policies, a subset of labour market policies. According to Clegg & Durazzi (2023), labour market policies are "government interventions that explicitly aim to shape the labour market risks confronting workers" (p. 5). This definition broadens the concepts of international entities like the European Union or the OECD, which had established datasets on a narrower set of policies (DG EMPL, 2018; OECD, 2022; Grubb & Keese, 2018; Grubb & Puymoyen, 2008). Inside this definition of labour market policies, we might include the following interventions: active labour market policies, passive labour market policies, employment protection legislation, collective bargaining policies, and minimum wage policies (OECD, 1990; Blanchard et al., 2014). Following seminal works in the literature, it is useful to understand active labour market policies in relation to the other labour market policies (Calmfors, 1994; OECD, 1990; DG EMPL, 1990; Rehn, 1989). Let us define what we mean by each subset of interventions paying greater attention to the active ones.

Firstly, active labour market policies (ALMPs)[4] are interventions that pursue the transition or retention of individuals to/in employment (Cronert, 2022; Cahuc et al., 2014; Auer et al., 2008)[5]. Note that after socioeconomic changes started at the 1970s, ALMPs also aimed at transitions from one to other employment considered

[4] They were originally called active manpower policies (Meidner, 1969).

[5] Cronert (2022) also considers the transition to education programs as the objective of ALMPs. We argue that it is more coherent to understand this as an intermediate objective and (stable) employment as the final objective.

better in terms of quality (Bonoli, 2007)[6]. Theorists have also posited that ALMPs might have secondary objectives like avoiding inflation or wage pressures (Erixon, 2010; Calmfors, 1994). The concept of ALMPs has conventionally included four tools: job search services, adult training, employment subsidies, and geographical mobility subsidies. We will see later that their differences are related both to the type of tool and to the channels used to impact the employment probability. In what follows we advance some ideas and dig into the historical roots of ALMPs.

Following Bonoli (2010; 2013) and Clegg and Durazzi (2023), the origin of the ALMPs was in Sweden during the 1950s with the Rehn-Meidner model. The tools included under the label of ALMPs were not new neither in the economic theory nor in public policy applications at that time (Weishaupt, 2011; Auer et al., 2008; Rehn, 1990). The novelty was the development of an economic theory that encompassed three policy fields under the same macroeconomic framework: fiscal policies, monetary policies, and labour market policies (Erixon, 2010). For recessions, Rehn (1985) explicitly opposed the inflationary expansive policy of generalized public demand with his anti-inflationary expansive policy through targeted public demand. On one side, the Rehn-Meidner theory promoted the *supply* of labour emphasizing the importance of mobility (Rehn, 1989). This concept was central in his theory and regarded both geographical and occupational mobility, whether vertical or horizontal. Rehn (1989) admitted that there were people with a low willingness to move, but he argued that even they would benefit from ALMPs. The idea was that those with a sufficient willingness to move will do it if costs are affordable and will leave their previous occupational/local labour markets less tight. On the other side, the Swedish economists highlighted the need of additionally increasing the *demand* of labour. Rehn (1985) defended employment subsidies both for public and private jobs, but also suggested using other fields of economic policy. For instance, he considered positive for this end to use investment funds, export credits, public house building, or public demand (Rehn, 1989).

The policy reception of these interventions has varied over time beyond Sweden. The OECD rapidly supported the ALMPs at its recommendations of 1964, 1966,

[6] This includes transitions from temporary to permanent employment, from part-time to full-time employment, or from mismatched to matched employment (whether vertical or horizontal).

1968, 1970, and 1976 (Alujas Ruiz, 2002)[7]. During the 1960s, Weishaupt (2011) describes how these reports seem to have influenced the policies of job search services and adult training passed in Denmark, Germany, Austria, United Kingdom, and Ireland. The unemployment rose in the 1970s and ALMPs were still promoted in European countries like Sweden or Germany, but changing the focus from adult training to employment subsidies (Bonoli, 2013). This reorientation was supported by studies like Baily & Tobin (1977), who highlighted the positive effect of employment subsidies, but adding cautionary messages on displacement and production effects. The 1980s were characterized by a greater divergence in the adoption of ALMPs across European countries, with the Nordic countries maintaining the focus on ALMPs and Anglophone countries drawing attention to the rest of labour market policies. Continental Europe invested more money in PLMPs than in ALMPs, usually reserving the latter for insured workers, in contrast to the Nordic strategy of expanding the labour force (Weishaupt, 2011). In the academic sphere, Calmfors (1994) took up the challenge of studying the interactions of ALMPs with other labour market policies. The 1990s and 2000s had this broader vision on all labour market policies, but now the EU took the role of principal advocator of ALMPs (Weishaupt, 2011). During the 2010s, we have seen in the European policy discussion the integration of the ALMPs inside larger policy packages sometimes called active social policy or social investment, in which there is space for work-life balance policies or active ageing (Bonoli, 2013; Hemerijck, 2017).

Secondly, passive labour market policies (PLMPs) are interventions aimed at alleviating the economic losses caused by unemployment or inactivity (Cahuc et al., 2014; Auer et al., 2008). Their tools are money transfers in the form of unemployment benefits, public-funded redundancy compensations, public-funded bankruptcy compensations, or early retirement benefits (DG EMPL, 2018; OECD, 2022). In short, if an ALMP is a trampoline towards employment, a PLMP is a life net for the fall in income[8]. The flagship of PLMPs is the unemployment insurance, a social insurance system that was first implemented by a public entity in Ghent in 1901 (Weishaupt, 2011). France was the first

[7] It is important to consider that Rehn was the Director of the OECD's Manpower and Social Affairs Directorate from 1962 to 1974 (Weishaupt, 2011).

[8] The same metaphor was used as subtitle in World Bank (2001): "From safety net to springboard".

country to execute this policy in all its territory in 1905. Then, unemployment insurances were sequentially appearing in Europe with some differences in design features, like the mandatory condition or the implementation through unions (Weishaupt, 2011).

Interrelating ALMPs and PLMPs, the literature has tended to conflate two conceptions of *active* labour market policies (Bonoli, 2011). The Swedish interpretation highlights that the *government* should be active by intervening the economy to correct labour market failures. The English interpretation emphasizes that the *benefit recipient* should be active by searching for a job while receiving public funds. We use ALMPs exclusively to regard the first interpretation and, following Clasen & Mascaro (2022), we leave the second meaning to the concept of activation or conditionality in the design of PLMPs.

Thirdly, employment protection legislation (EPL) is a set of public policies that regulate the processes of hirings and dismissals. For this reason, they are specially related to the so-called external flexibility of firms. Regulation materializes as unstructured data in documents gathering decisions at many levels of policymaking, so the characterization of the design of these policies is a complex task. To do it, the OECD created the EPL Database (OECD, 2020). These regulations usually set different rules for hirings with open-ended and temporary contracts, and for individual and collective dismissals. Some of the aspects regulated for the individual dismissals of workers with open-ended contract are "procedural requirements before notice can be given, notice period and severance pay, the regulatory framework for unfair dismissals and enforcement of unfair dismissal regulation" (p. 178). Regarding the hiring of workers with temporary contracts[9], some of the items delimited in these policies are the valid reasons to use these contracts, the maximum number of consecutive temporary contracts, the maximum cumulated duration of this sequence of contracts, or the requirement of authorizations or declarations of these contracts.

Fourthly, collective bargaining policies (CBP) are a set of public policies that delimit the processes of setting wage and working conditions agreed between trade unions and employer organizations. Collective bargaining agreements are private policies, because decision-makers are private agents. However, public authorities

[9] Inside the category of temporary contracts, the OECD (2020) includes both the fixed-term labour contracts and the temporary agency work contracts.

may intervene in these pacts through regulations, that may have different features in respect to their goals and means[10]. Let us review examples of these rules to influence the collective bargaining coverage, control, and centralization (Visser, 2024; Visser, 2021). Regarding coverage, extension policies force the application of collective agreements to all workers in its jurisdiction, not only to members of signatory unions (Visser, 2016). This policy may take different legal forms, like a provision of a law, an administrative order, or judicial decisions (Visser, 2024). Moreover, temporal validity policies[11] delimit the duration of collective agreements by establishing a minimum/maximum duration as a default or mandatory option. Notice that the minimum duration may even be indefinite by default, as in Portugal until 2003, where the collective agreement could only be terminated if the parties explicitly settle it (Visser, 2016).

There are other CBP that may facilitate the enforcement of the agreements through control. Policies on works councils may make their presence in each establishment something mandatory, but other policies may also force the union representation in the workplace as separate from these work councils. Additionally, CBP may also regulate the conflicts of rights by referring grievances through an arbitration procedure established in the law. CBP may also interfere on the degree of centralization of collective bargaining. Favourability policies impose a hierarchy of levels of bargaining in which a certain level is considered superior in application. For instance, sectorial agreements may take priority on enterprise agreements, or the opposite. The favourability rule is usually understood in a positive way for the worker, meaning that even under sectorial favourability, an enterprise agreement may be applied rather than the sectorial *if* the former improves the employment conditions of the latter (Visser, 2024).

Fifthly, minimum wage policies (MWP) are the alternative path took by some governments to delimit the processes of setting wages. These policies also use regulations as tool but are substantive instead of procedural because government directly establish the minimum wages without requiring the approval of social

[10] For the analysis of policies that use regulation tools, it is useful to consider an approach to codify the documents of a policy design known as Institutional Grammar (Frantz & Siddiki, 2022). After splitting the documents in units of analysis (usually statements), one of the dimensions proposed by this approach is the prescriptive operator of the unit. This variable identifies if a provision of a regulation requires (must), allows (may), or forbids (must not) a certain action. Therefore, the policies presented below are just particular cases of certain prescriptive operators.

[11] They are also known as ultra-activity, after-effects, or temporal extension policies.

actors. Thereby, to difference these rules from the wage floors set in collective bargaining, they are sometimes called statutory minimum wages. Picot (2023) argues that the main design features of MWP are: the coverage (national or sectorial), the level, the process to adjust such level, their differentiation in function of the worker category, and the enforcement rules.

Now that we have briefly presented all labour market policies, we may concentrate on the two sides of the ALMPs studied in our doctoral thesis. We try to bring evidence about the effectiveness of different design features (on the means side) and about the selection of different target groups (on the goals side). Let us review what have been the conventional solutions for configuration of tools and target groups in ALMPs.

1.2.1. Tools of ALMP

The tools or instruments included in policy designs to provoke behavioural changes on the target individuals are usually the fundamentum of many ALMP classifications. There have been different typologies proposed by international organizations (DG EMPL, 2018; OECD, 2022), researchers from economics (Brown & Koettl, 2015; Rehn, 1985) or from social policy (Bonoli, 2013). We argue that there are many common points between these classifications, so we are going to take ideas from different sources. We use a hierarchical classification of ALMPs that follows the two dimensions of goals and means of public policies. The first level of this classification is available at Table 1. Initially, we are going to explain each dimension of classification. Then, we deep into each cell to describe how each tool may contribute to each goal.

Table 1 – First-level classification of ALMPs

	Matching	*Productivity*
Service provision	Job search services	Adult training
Money transfers	Geographical mobility subsidies	Employment subsidies

Note: The columns refer to the goal and the rows to the mean.

Regarding the goals dimension, Crépon & van den Berg (2016) propose an interesting way of classifying ALMPs according to their main objective[12]. There are programs aimed at facilitating the *matching* process between employers and workers. Inside this group we may place the job search services and the geographical mobility subsidies. In a sense, these policies try to modify the supply of labour by acting first on the behaviours, preferences or beliefs of workers (Rehn, 1985). There are other programs directed at increasing the *productivity* of the workers. On one side, adult training attacks the numerator of the productivity function by increasing the productive skills of workers (the output). On the other side, employment subsidies influence the denominator of the productivity function by decreasing the labour costs (the input). Complementary, these programs are aimed at changing the demand of labour by intervening first on the behaviours and beliefs of employers (Rehn, 1985). In sum, matching interventions immediately reduce the costs of job search to the jobseeker, whereas productivity interventions immediately decrease the costs of the employer.

Regarding the means dimension, we think that it is useful to depart from the tool classification developed by political scientists across different policy fields (Howlett, 2018). In the previous section, we summarized it in three clusters of tools: goods/services provision, money transfer, and regulation. We saw before that regulation is the tool of EPL and CB, leaving the other remaining clusters to ALMPs and PLMPs. Thereby, ALMPs work either through service provision or through money transfers. A similar distinction has been proposed by Weishaupt (2011) and OECD (1990) by distinguishing financial and non-financial interventions. As we can see in Table 1, if we follow the original proposal of Rehn (1989) that gave remarkable importance to mobility, we can place ALMPs in the four cells resulting from combining the means (services or money) and the goals (matching or productivity).

The classification proposed in Table 1 may be further detailed either decomposing the design features of the intervention (on the means side) or looking at the intermediate variables targeted by each intervention (the goals side). Let us introduce now each tool and the theories that justify public funding of these interventions. The research chapters devote a deeper literature review on the uses

[12] This idea is also in Kluve (2010), but without explicitly linking employment subsidies to productivity.

and effects of job search services and adult training, since they have been the tools studied in this doctoral dissertation.

Job search services. The literature has usually proposed a miscellaneous concept that incorporates job search assistance, monitoring, and sanctions (Card et al., 2018); enhanced services, which include assistance and sanctions (Vooren et al., 2018); employment services (Kluve et al., 2019); labour market intermediation services (Escudero et al., 2018); or just job search assistance (Filges et al., 2018). Our suggestion is using "job search services", a concept that encompasses both assistance and intermediation. We argue that it is useful to separate job search assistance and job search intermediation because they intervene on the job search through two different channels[13]: *improving* the job search made by the worker or *delegating* such job search in an intermediary. Assistance per se does not remove the job search as a task of the jobseeker, whereas intermediation substitutes the job search of the participant by assigning the task to a caseworker. Both tools may be mixed, e.g. having both better job search developed by the worker and job search performed by a third-party, but they represent two different strategies to cope with the problem. Thereby, public or private employment offices might use one or both strategies.

Job search intermediation has sometimes been called vacancy referral (Bollens & Cockx, 2017; Huuskonen, 2023; Jeremy, 2023), employer intermediation (Brown & Koettl, 2015), job placement (Cahuc et al., 2014), or simply labour market intermediation (Clegg & Durazzi, 2023; Escudero et al., 2019; Dammert et al., 2015). Another way of thinking on job search services is as a cumulative continuum of services sorted from less to more proactiveness by part of the service provider[14]. First, the less proactive service is what Autor (2009) calls a basic job board, which is simply an offline or online platform in which many employers pool their vacancies and workers register their profiles as jobseekers. Second, the next degree of proactiveness might be what we call a closed vacancy referral. This involves an evaluation of the suitability of several vacancies for a certain jobseeker (if the service is for a worker). In the setup of ALMPs, under this service the

[13] Inside job search services, besides job search assistance and intermediation, some researchers introduce the job search requirements (Cahuc et al., 2014; Kluve, 2010). These requirements, whether soft (monitoring) or hard (sanctioning), are always linked to PLMPs. For this reason, as we argued before, we place job search requirements (conditionality) as a tool feature of PLMPs.
[14] Inversely, we can think them as a continuum from more to less automatable.

worker fixes the value of some parameters of the search like the occupation, the industry, or the location. If these values match those of more than one vacancy, the intermediary may assess how the search parameters set by the employer align with those of a particular jobseeker. This deals with the other part of the intermediation, because the provider may evaluate several jobseekers for a certain vacancy (giving a service for an employer). The following degree of proactiveness involves interventions aimed at modifying the behaviors or preferences of the jobseeker, which we refer to as job search assistance.

Job search assistance includes interventions that have been deeply studied by psychologists. Liu et al. (2014) propose two subsets of tools depending on the intermediate outcomes that they target (job search skills or jobseeker's motivation). The first group may include actions like training on job search skills to approach vacancies via internet or networking, but also training on self-presentation abilities. These interventions target the quality of the job search. The second group focuses on psychological elements like self-efficacy, proactivity, goal setting, social support or stress management. These interventions are aimed at the quantity of the job search. Junquera (2024a), based on Crépon & van den Berg (2016), has proposed a third subset to include those tools directed at a third intermediate outcome: the job search targets. The job search targets encompass the occupations, sectors, and geographical areas of interest. The interventions in this group are mainly based on the provision of costly or difficult-to-access information on labour or product markets. They foster a compromise between the supply and the demand of labour. We may think them as a redirection of the search or as an *open* vacancy referral. The main idea is that the employment office tries to change some of the key search parameters to have an earlier or more socially efficient match.

The existence of employment offices might be explained by an incomplete information problem. Collecting information on vacancies is costly for jobseekers. Equivalently, gathering information on jobseekers is costly for employers. In this situation, there is a place for firms in the form of employment offices that facilitate the matching process through assuming the high fixed cost of collecting employers with vacancies and jobseekers (Cahuc et al., 2014).

But why should the government fund a public employment office? Lee (2009) argues that the public provision of these services is justified by an asymmetric

information problem between jobseekers and private providers of intermediation. Departing from true cases of abuse of private providers towards jobseekers in USA[15], Lee (2009) maintains that if workers cannot differentiate between low-quality (abusive) and high-quality (non-abusive) providers, the latter have an incentive to offer low-quality services. This might be a reasonable assumption due to the lower educational attainment of clients of these services and to the problems associated to incomplete contracts (Lee, 2009; Cahuc et al., 2014). The consequence is that only low-quality providers remain in the market, a situation that is known as adverse selection. The introduction of public providers of intermediation may solve the problem by crowding out the low-quality private providers. The argument is that uninformed jobseekers might use public employment services (PES) without cost and without risk of abuse. To face this situation, low-quality private intermediaries may either exit the market or offer a credible high-quality service[16].

Adult training. This subset of ALMPs has also been related to multiple names such as retraining (Vooren et al., 2019; Kruppe & Lang, 2020), labour market training (Filges et al., 2018; Andrén & Gustaffson, 2004), continuous training (Lechner, 2000), vocational training (Hirshleifer et al., 2016; Andrén & Andrén, 2006), or simply training (Card et al., 2018; Escudero et al., 2018; Kluve, 2010). They do not always refer to the exact same intervention, because the concept sometimes highlights a specific tool feature. We support the use of "adult training" to be more specific with the interventions usually framed under the label of ALMP and to be as general as possible to include different configurations of their design. Inside the European Union, CEDEFOP (2014) makes an interesting distinction between *initial* education/training and *adult* education/training[17]. This recognizes that vocational training may be offered inside the educational system (initial) or outside it (adult). We argue that the idea of adult training is aligned with two design features that were in the origins of ALMPs in Sweden.

---

[15] Some examples of abuse were the following: providers of intermediation not making effort after jobseekers paid the fee, referral of vacancies unfeasible for the jobseeker due to distance reasons, collusion with employers for dismissals…

[16] We must recognize that the more the public provider offers a credible high-quality service, the more it pushes private providers to increase their quality to compete. However, Larsen & Vesan (2011) remind that high quality of job search intermediation in PES is not something trivial.

[17] The ISCED 2011 also has a similar distinction between adult and initial education (UNESCO Institute of Statistics, 2012).

One of the interventions promoted during the origin of ALMPs in Sweden were called quick retrainings (Bonoli, 2010). This concept highlights two important features of these interventions: its non-initial character and its short duration. On the non-initial character, Rehn (1985) argued that labour market training programs had always being separated from the educational system. He defended the importance of "generous help to workers to leave imperilled fields (by providing retraining or moving allowances) and take up jobs in more prosperous fields and areas" (Rehn, 1985, p. 66). On the short duration, notice what Rehn (1985) calls the original principle of the system: "the ability to make rapid improvisations in courses to reflect changes in the local labour market situation and to utilize as teachers anyone with experience in a given occupation, without much reverence for formal certifications, etc." (p. 75). This idea crashes with the rigid calendars and procedures of the educational system. Moreover, it indicates that these training courses are generally shorter than those offered under the educational legislation.

These two characteristics are generally shared by all training interventions under systems of ALMPs, but there are also substantial differences in other features of the programs. Kemper et al. (2022) gathered some variables that may be called features of adult training like the learning model (classroom-based vs. work-based), the intensity (usually proxied through duration), the degree of formality (including certification and regulation of curriculums), or the economic support to beneficiaries (through cash or services like childcare). Other studies have focused on the target groups mentioned in the policy goals, like unemployed workers or employed workers (the so-called continuous training). We have specially concentrated on the first two design features. Regarding the learning model, the OECD (2018) defines work-based training as "learning that takes place through some combination of observing, undertaking and reflecting on productive work in real workplaces" (p. 29). This learning strategy is sometimes called internships when there is not a payment to the trainee and apprenticeships when there is (OECD, 2023d). Concerning the intensity, the extension of the learning is difficult to observe, but we can use duration as a proxy of the quantity of hours devoted to such learning or the amount of money paid to the teacher as a proxy of the quality of such learning process (Escudero et al., 2019; Martin, 2015).

Economists have justified the public funding of adult training both when there is imperfect or perfect competition in labour markets (Cahuc et al., 2014). It is

useful to distinguish between on-the-job training (for employees) and out-of-the-job training (for unemployed workers). The type of ALMP that we have called adult training is closer to the Becker's definition of general training, since it is not firm-specific (Leuven, 2005). Regarding out-of-the-job training, Cahuc et al. (2014) argues that informational asymmetries may produce several social inefficient individual decisions on general training. On one side, an overeducation of the worker due to the difficulties to signal her productivity. On the other side, undereducation due to imperfect credit markets (also argued in Boeri & van Ours, 2008), transaction costs, or the difficulty to signal her productivity if she was trained. Concerning on-the-job training, Boeri and van Ours (2008) maintain that public funding of these programs is only justified when labour markets are not monopsonistic. When the situation is closer to perfect competition due to lack of market power and lack of search frictions, firms have no incentives to invest on training because their workers can easily change to another company. We will focus on adult training for unemployed workers.

From the social policy literature, the participation of individuals in adult training programs has been promoted due to the "(changing) skill requirements on the supply side of the economy" (Hemerijck et al., 2023, p. 485). In a sense, this also compels the state to fund adult training, since those now with mature age took their initial training under a very different socioeconomic structure. For example, changes like the tertiarization of the economy were promoted by the state through free trade agreements. Thereby, the state varied the "rules of the game" after individuals invested on certain education decisions now obsolete, so its intervention may be understood as a compensation.

Employment subsidies. Under this label, like Crépon & van den Berg (2016) and Kluve et al. (2019), we place interventions that subsidize labour costs either in the private or the public sector. The programs aimed at the private sector have received names like subsidized labour (Vooren et al., 2019), subsidized private sector employment (Card et al., 2018), private sector incentives (Escudero et al., 2018; Filges et al., 2018), wage subsidies (Jaenichen & Stephan, 2011; García-Pérez & Rebollo-Sanz, 2009; Katz, 1998), or marginal employment premiums (Rehn, 1990). The programs directed towards the public sector have been called public sector employment (Vooren et al., 2019; Eichler & Lechner, 2002), subsidized public sector employment (Card et al., 2018), public works (Escudero et al., 2018), or direct employment programs in the public sector (Filges et al.,

2018). The concept of "employment subsidies" is useful to highlight the common base of all these programs that differentiates them from adult training: they fund labour costs through money transfers without doing efforts to guarantee a minimally formalized training experience. All these interventions incentivize hiring by offering an increase in the productivity of the organization.

That said, a second level classification is useful because it clarifies the different mechanisms (intermediate variables) targeted by each type of employment subsidy. On one side, employment subsidized at the private sector may influence the employment probability after the subsidy due to a second mechanism: signalling (Crépon & van den Berg, 2016). If employers are reluctant to hire the targeted individuals due to the lack of information on its productivity or statistical discrimination, these interventions might subsidize part of this risk by allowing employers to assess such productivity in the workplace. The downside is that, because hirings precisely depend on the risk perceptions of private employers, recruitment could take more time to happen than in public jobs. On the other side, employment subsidized at the public sector may not have an effect through the signalling channel due to the highly bureaucratized character of many public administrations. In these cases, hiring is widely standardized through a set of pre-specified variables, so unobservable factors like productivity are usually not considered directly in the recruitment process.

Different studies on ALMPs have indicated how wide is the space of design features linked to these interventions (Cronert, 2019; Crépon & van den Berg, 2016; Katz, 1998; Rehn, 1990). Some of these variables are the duration, the subsidy amount (as a proportion of the wage or a lump sum), the transfer channel (e.g. cash, a tax decrease, or a social security contribution reduction), and the characteristics of the subsidized agent (the employer). Cronert (2019) analysed 3,464 versions of programs of employment subsidies implemented in Europe from 1998 to 2013. He proposed two latent dimensions to group these subsidies while collecting 79 % of the variance in several variables of the design: how outsider might be considered the target group and how structural/cyclical might be considered the labour market challenges faced by the subsidy.

The justification of the public funding of employment subsidies has usually been framed in a macroeconomic framework. It is useful to depart from the existence of equilibrium unemployment due to the existence of efficiency wages, unions more

focused on insiders than outsiders, and/or the moral hazard problem of unemployment benefits (Layard et al., 2005). To cope with this equilibrium unemployment, governments may implement expansionary demand policies, but this may generate inflation problems. Baily & Tobin (1977) suggest that an employment subsidy targeted towards unemployed workers increases the aggregate demand and "contribute less to wage pressure than a corresponding increase in regular employment" (Calmfors, 1994, p. 24). The reason is that increasing public demand *without* fostering the hiring incentives on certain groups of workers might generate a competition of firms across employed workers, which would raise wages.

Geographical mobility subsidies. This category of interventions has not been included in contemporaneous literature reviews on ALMPs, probably due to the scarce number of evaluations of these programs (Caliendo et al., 2023). This is interesting due to the importance of geographical mobility subsidies in the Rehn's theory and in the Swedish policy of the 1950s and 1960s (Erixon, 2010). In the proposal of Rehn (1990), to decrease the unemployment rates, the subsidized agent is not necessarily an unemployed worker. Geographical mobility subsidies may push employees in an overheated territory to move to the incentivized territory, leaving vacancies for those unemployed workers that are not willing to change their residence.

We may create a second level classification for these subsidies by distinguishing two possible channels to influence the employment probability: subsidizing commuting or subsidizing residential changes. Inside these categories we can also distinguish some design features. For commuting, Potter et al. (2006) describes how some developed countries have experimented with subsidies on the costs of public or private transports, whether by car, bus, weekly plane, or bicycle. For residential changes, Caliendo et al. (2023) reports that Germany designed different incentives for the cases in which (a) the costs of getting a new job included to maintain two households or (b) this new employment required a full relocation with higher transportation costs. In any case, policymakers need to make a decision on other design features like the subsidy amount (whether as proportion or as lump-sum pay), the duration, or the transfer channel (e.g. through the tax system or through cash).

The existence of public geographical mobility subsidies may be grounded on the existence of structural unemployment due to a wage-setting procedure that depends on the unemployment rate of its territory *and* the unemployment rate of a leading territory in the same region. For example, let us think about a provincial sectorial collective agreement determining wages in the municipality of Barcelona and also in a small municipality of the province of Barcelona. If wages in the leading local labour market cannot decrease more, subsidizing mobility to less tight local labour markets is justified due to the negative externalities of unemployment or due to other externalities like traffic congestion. The funding of these ALMPs could be done by increasing taxes on the leading territories according to a Pigovian logic (Layard et al., 2005).

Lastly, we may think that the four tools usually understood as ALMPs have two modalities: aimed at employment or aimed at self-employment. Job search services for self-employment may offer information on possible product markets to enter or may delegate tasks related to firm creation. Thus, labour market intermediation might be understood as the sum of job search intermediation (a service for workers) and job offers intermediation (a service for employers). Adult training programs for self-employment would train people on legal and economic aspects to manage a company. Employment subsidies for self-employment are the so-called start-up incentives, which may not only fund labour costs, but also capital costs (Román et al., 2013). In relation to geographical mobility subsidies, some governments have offered housing grants in a more general framework than labour market policies, so both employees and people who is self-employed are eligible (van Dijk, 2019).

1.2.2. Target groups of ALMP

Another decision that policy designers must make is the target groups of the policy goals. Since ALMPs are implemented to face not only transitions from unemployment to employment, they have not been only directed at unemployed workers. Cronert (2022) took all the ALMPs except job search services executed in the EU plus Norway from 1998 to 2013 and concluded that 87 % were targeted towards the unemployed, 38 % towards the inactive, and 23 % towards the

employed[18]. This is a one-dimensional approach to define target groups, but nothing precludes to classify individuals into multi-dimensional target groups (e.g. low-skilled unemployed, high-skilled unemployed, low-skilled inactive…). Thereby, the number of target groups that may be made up is infinite, so policy designers usually deal with a shorter list of candidate target groups. Beyond philosophical considerations on which group to attend, the literature has attempted to put forward two main paths for selecting target groups in a data-driven manner[19].

Firstly, one reason to attend the selection of target groups is that interventions may have heterogeneous effects in function of the characteristics of the participants (Kluve et al., 2019). In potential outcomes terms, conditional average treatment effects on the treated for each program $D = d$ are usually estimated for each candidate target group $X = x$. Formally, this estimand is $\mathbb{E}[Y_i(D = d)|D = d, X = x] - \mathbb{E}[Y_i(D = 0)|D = d, X = x]$. Such counterfactual view of ALMPs allows to consider the possible deadweight effects (European Commision, 2017). Notice that it is not necessary to restrict the analysis to the treated population, but also consider the displacement effects on the untreated population (Cheung et al., forthcoming; Crépon & van den Berg, 2016). The classic approach to heterogeneity is called subgroup analysis, because groups are defined in advance and then potential outcomes are estimated (Yang et al., 2021). The contemporaneous approach to heterogeneity is more prone to define groups *after* potential outcomes are estimated using techniques like causal forests (Imai & Ratkovic, 2013; Wager & Athey, 2018).

Secondly, when treatment effects have not been credibly estimated for different target groups, a more modest approach is to allocate treatments according to the unemployment probability of an individual in the future under no interventions. In potential outcome terms, the estimand of this approach is $\mathbb{E}[Y_{iT=t}(D = 0)]$, where t is usually one year later in labour market policy applications. If the previous strategy allocates resources according to how beneficial a treatment is for a group, this strategy does it according to how severe is her diagnosis. This procedure is usually called jobseeker profiling (Desiere et al., 2019; Körtner & Bonoli, 2023). While the previous strategy is focused on effect estimation, this

[18] The percentages do not sum up to 100 because a program may be targeted to two or three of these groups, i.e. they are not categories of a unique variable.

[19] Kuppler et al. (2022) is an interesting outlier in the literature, because it combines philosophical debates on distributive justice with statistical procedures for choosing target groups.

one is concentrated on prediction. In practice, jobseeker profiling usually joins the pre-determined and the post-hoc grouping of individuals. To evaluate the predictive performance, it compares the observed with the predicted value using a variable whose categories are initially defined. To deploy the model, a PES may generate new groups from the predicted score attending its specific goals and costs (Kuppler et al., 2022).

Two additional dimensions may be considered when deciding targeting: stigmatization and inflation. On one side, Calmfors (1994) considered targeting as one of the crucial design features and introduced a warning: ALMPs should not be targeted in a way that give them bad reputation, because if this is the case employers may use them as screening devices. For this reason, he recommended an anticipation of future long-term unemployment by directing programs towards medium-term unemployed. On the other side, Rehn (1990) proposed selecting the target groups taking into account the inflation consequences. Specifically, he argued that focusing ALMPs on sectors with high product demand might relief the inflationary pressures derived from an insufficient product supply. This would avoid confronting the empty vacancies with a wage growth and would allow to decrease the unemployment rate.

2. Justification

This doctoral thesis analyses labour market policies implemented in two countries using insights from two disciplines. In the first part, we contextualize and justify the pertinence of analysing cases of studies from both countries. In the second part, we justify the relevance of writing a dissertation from the perspective of two social sciences.

2.1. Spain and Italy

We analyse three policies related to the labour markets of two countries, Spain and Italy, during the period 2015-2022. Specifically, we focus on two regions of these states: Catalonia and Veneto. To contextualize these countries, we are going to review for the period of interest: (1) the main economic and social features of these territories with their average values, and (2) the main features of their labour market policies. Then, we are going to present three reasons that justify

our dissertation in relation with these socioeconomic and policy configurations. All the following data come from Eurostat (2024a; 2024b; 2024c), except the statistics on collective bargaining, which are estimated in OECD (2023d) and Visser (2024).

Spain has 46.9 million inhabitants and a GDP per capita of 25,250 €, below the EU-27 average of 30,689 €. The percentage of the population aged 15-64 that is in the labour force is 74 % and the unemployment rate has ranged from 22.1 % in 2015 to 13 % in 2022. At the regional level, Catalonia has 7.5 million inhabitants and a GDP per capita of 29,750 €. Its unemployment rate has gone from 18.6 % in 2015 to 10 % in 2022. Regarding the income in Spain, the inequality of its distribution measured through the ratio of the highest versus the lowest quintile is 6.2, which is above the EU average (5.01). Also related to income, its collective bargaining coverage rate is 73 % and the union density is 13 %.

Italy has 60 million inhabitants and a GDP per capita of 29,539 €, closer to the EU-27 average mentioned before. The labour force participation rate among those aged 15-64 is 65 % and the unemployment rate has gone from 12 % in 2015 to 8.1 % in 2022. At the regional level, Veneto has 4.9 million inhabitants and a GDP per capita of 33,238 €. The unemployment rate has ranged from 7.1 % in 2015 to 4.2 % in 2022. Concerning income in Italy, the inequality of its distribution is similar to the one found in Spain with a value of 5.92, distant from Nordic countries like Denmark (4.05) or Continental countries like Germany (4.76). Such distribution is conditioned by a collective bargaining coverage rate of 100 % and a union density rate of 33 %.

Let us make now a brief review on the state of the different labour market policies at these two countries during the years of interest.

Active labour market policies. In Spain, ALMPs are designed and implemented at the three levels of government (state, region, and municipality), but most decisions are made at the regional level after the decentralization process ended in 2011 (Cueto & Suárez, 2015; Rojas, 2012). This has provoked that some regions like Catalonia established a quasi-market of ALMP providers for services like adult training. An important source of funding comes from transfers from the state to the regional level according to indicators of implementation, which are far from effect estimates of the programs (Cueto & Suárez, 2015; Cebrián López et al., 2021). As we see in Table 2, in Spain, the total expenditure per person

wanting to work in all ALMPs except job search services is 1,477 currency units at purchasing power standards (DG EMPL, 2020). Although it is not very far from the European average, it is still well below Nordic countries like Denmark or Sweden. If we look at the distribution of the expenditure across tools, we see that in Spain it is much more prevalent the use of employment subsidies than adult training. Something similar happens with the expenditure in job search services, being close to the EU average but having around one third of the resources devoted in Nordic countries.

The current system of ALMPs in Italy has been substantially delimited by the Jobs Act reform (2014-2016) of the Renzi Cabinet (OECD, 2019b). The policy governance is also multi-level and unequally shared between the central and the regional governments, giving more weight to the regions. Both in Italy and in Spain, the state level determines the minimum set of programs that every region should offer. Regarding funding, transfers from the state to the regions started as in Spain through yearly agreements during 2016 and 2017 but changed to a permanent transfer from 2018 (OECD, 2019b). Quasi-markets for service provision were promoted throughout the country with a first phase of jobseeker profiling and a second phase of participation in ALMPs through vouchers (Scarano, 2019). Table 2 shows that the actual expenditure of Italy in adult training or subsidies (either for employment or mobility) is very low in comparative terms. The 972 currency units per person wanting to work are very distant from the EU average. The distribution across tools resembles Spain, since it gives more weight to subsidies than to adult training. The spending on job search services, 0.002 % of the GDP, is even more negligible. There is a stark difference with the country with the largest expenditure, Germany, which invests 0.219 % of the GDP in this tool.

Table 2 – Country expenditure in ALMPs. Data for 2018 in purchasing power standards for members of the European Union plus Norway.

Country	*Total ALMP, except JSS*	*Adult training*	*Private employment subsidies*	*Public employment subsidies*	*Self-employment subsidies*	*Job search services (JSS)*
DK	8,269.2	17.65	82.35	0	0	0.195
LU	5,660.8	32.06	65.44	2.47	0.03	0.032
SE	5,600	11.9	87.45	0	0.65	0.181
FI	3,698	47.82	25.43	24.68	2.07	0.088
BE	3,466.6	31.21	60.63	8.03	0.13	0.100
AT	3,167.1	73.47	16.44	9.18	0.91	0.106
HU	3,062.3	4.81	21.99	69.9	3.3	0.044
NL	2,904	16.14	83.86	0	0	0.078
NO	2,652.6	33.81	65.74	0	0.44	0.094
CZ	2,508.2	1.92	79.26	18.32	0.51	0.026
IE	2,255.4	39.8	12.34	47.86	0	0.008
DE	2,216.3	71.09	18.13	7.27	3.51	0.219
EU	2,219.4	29.16	53.95	13.7	3.19	0.061
ES	1,477.3	20.6	34.77	21.38	23.26	0.067
EE	1,328.9	27.49	70.07	0.08	2.37	0.101
PL	1,295.2	1.91	81.83	4.18	12.09	0.027
HR	1,221.6	25.46	24.72	27.37	22.45	0.053
LT	1,197.8	33.46	66.53	0	0	0.020
PT	1,101.2	52.86	37.07	8.91	1.17	0.036
SK	973.8	13.62	76.48	4.62	5.28	0.002
IT	972.8	30.01	69.28	0.61	0.1	0.002
SI	865.9	32	27.44	37.87	2.7	0.042
MT	562.7	7.82	90.74	1.44	0	0.022
BG	440.7	28.59	27.18	44.04	0.16	0.000
LV	376.4	52.15	29.46	16.82	1.57	0.023
CY	280.6	36.89	63.01	0	0.11	0.016
RO	149	13.56	85.17	1.14	0.2	0.000

Note: The data come from DG EMPL (2020), which does not have available data for France, Greece, and United Kingdom. The first column shows the total

expenditure per person willing to work (those unemployed plus inactive people who want to work but are not searching or immediately available to work) in absolute terms at purchasing power standards. The rest of columns show row proportions. This is the last available analysis including data from Italy. In the EU database, geographical mobility subsidies are included in the private employment subsidies category (4). The last column shows data on the country expenditure in 2018 as percentage of its GDP and comes from DG EMPL (2024).

Passive labour market policies. The generosity of the system of PLMPs in Spain and Italy may be approximated through the replacement rates exposed in Table 3. The proportion of previous in-work income maintained after unemployment are always above the OECD average even after five years in Spain. The situation in Italy is similar until the first year. From the second year, the amount of income provided by PLMPs decreases substantially in relation to previous earnings. The conditionality of the PLMPs is especially high in the sanctioning dimension in the two countries both for the insurance and for the means-tested benefits. Specifically, the unemployment insurances of Spain and Italy are placed in the positions 6 and 7 out of 35 member states of the OECD for this dimension of strictness. These indices collect the conditionality according to program designs but looking at indicators of policy implementation in Spain we estimate that only 1 % of the unemployment benefits managed by the central PES were linked to a sanction (Ministry of Labour of the Government of Spain, 2024). Both countries are in the bottom of the classification of strictness of job search requirements procedures for this benefit (Immervoll et al., 2020).

Employment protection legislation. Spain has an intermediate degree of strictness on the dismissal of permanent workers, being placed at the 15th position out of 37 developed countries studied in OECD (2020). Italy has a high strictness on the dismissal of permanent workers in relative terms, since it ranks at position 5 out of the 37 states. In this country, some elements of the EPL for these workers are established through collective bargaining, like the notice or the trial periods. According to the data from OECD (2020), "in Italy, sectoral agreements generally provide for shorter trial periods and longer notice periods" (p. 192). On the degree of strictness of dismissals with temporary contracts, both countries are in the intermediate group one next to the other. Their indexes lie very close to the

average value for OECD members, with Spain just below Germany in the ranking (OECD, 2021)[20]. Notice that the variability between countries in the regulatory strictness to fire a permanent worker (CV = 0.180) is lower than the variability found to fire a temporary worker (CV = 0.457)[21]. Lastly, Spain and Italy are countries located in the group of high strictness of regulation for hiring temporary workers, since they are at the 5th and 3rd positions in this ranking (OECD, 2020).

Table 3 – Replacement rates of PLMPs in Spain, Italy, and OECD. Averages for [2015, 2022].

	After 2 months	*After 6 months*	*After 1 year*	*After 2 years*	*After 5 years*
Spain	77.75	77.75	57.125	57.125	33
Italy	73.375	70.625	60.5	30.875	15
OECD	64.25	56.5	44.125	35.5	28.375

Note: The data come from OECD (2024). The values are estimated for a household of a single person without children that had 67 % of the average wage as previous in-work incomes. The columns indicate the time elapsed since the entrance in the unemployment status.

Collective bargaining policies. Let us review the configurations in the two countries using data from OECD (2023d)[22]. Both Spain and Italy have extension policies, in Spain through *erga omnes* collective agreements and in Italy through the case law on pay scales. The temporal validity of collective agreements is of one year after expiration in Spain and if there is not new agreement, the higher-level

[20] This classification is made through the index of termination of fixed-term contracts, which is an average that gives equal weight to the indicator of protection against terminating fixed-term contracts before the end of the contract and the indicator of protection against terminating fixed-term contracts at the end of the contract (OECD, 2021). The clustering in three groups (high, intermediate, and low regulatory protection) follows OECD (2020). It considers as the high group the 10 countries at the top of the ranking, as the low group the 10 countries at the bottom, and the rest is assigned to the intermediate protection level.

[21] CV denotes the coefficient of variation.

[22] We used the values of the last year available (2019).

document applies (if any)[23]. We do not have evidence on the presence of such mandatory public policy for Italy (Visser, 2016). To influence the enforcement, the union workplace representation is mandatory in Italy, but in Spain is only guaranteed if they have negotiated a collective agreement. The situation is the opposite with respect to policies on arbitration procedures. Spain forces collective agreements to include a specific procedure for handling grievances, whereas Italy does not. Regarding centralization, the favourability policy of Spain gives priority to enterprise over sectorial agreements[24]. However, workers in most sectors were not finally covered by enterprise agreements, possibly because "most new collective agreements were signed by small or medium-sized firms, with low coverage" (Rocha, 2018, p. 257). The policy in Italy sets sectorial favourability as standard but allows unions and employers to deviate from it if sectorial collective agreements allow these deviances (Visser, 2024).

Minimum wage policies. Spain has maintained the following design features along the years of interest: national coverage, no differentiation between workers, and incorporation of some interventions to enforce the policy (Eurofound, 2023). However, the level of the statutory minimum wage has experienced remarkable changes, growing from the 655 € in 2016 to the 1,000 € in 2022 (Gorjón et al., 2022). Picot (2023) argues that the process of minimum wage updating cannot be characterized by an indexation or by the involvement of social actors, but by a "full government discretion" (p. 242). On the contrary, Italy does not have a statutory minimum wage, as minimum wages are set by social partners in sectoral collective agreements.

We have already seen that there are remarkable similarities between Spain and Italy. This is something that is usually recognized even more broadly in the literature on varieties of capitalism, grouping both countries as mixed market economies (Molina & Rhodes, 2007). Beyond the similarities, we argue that there are three main reasons to develop a doctoral thesis focused on both countries: the weak tradition of evaluation in these contexts, the high unemployment in some of

[23] This changed in 2022, because the Labour Reform Act established the indefinite temporal validity of collective agreements (Rojo Torrecilla, 2022).

[24] This also changed in 2022 with the Labour Reform Act, because it removed the enterprise favourability on wages (Rojo Torrecilla, 2022).

their regions, and their low expenditure in ALMPs. These three grounds correspond to a scientific reason, a societal reason, and an economic reason.

The first reason is the low tradition of evaluation in Spain and Italy. Looking at the most citated survey of ALMP evaluations, only 20 out of the 857 estimates came from a program implemented in Italy or Spain (Card et al., 2018). Specifically, these 20 estimates were the sum of 7 studies with Spanish data and 2 studies with Italian data. Compare these values with the 40 studies found for Germany or the 21 studies did for Sweden. This would not be a problem if there were not interactions between the effect of ALMPs and the effect of other labour market policies implemented across each state. However, authors like Boeri & van Ours (2008) or Blanchard et al. (2014) have made a case on the existence of policy interactions. It might be the case that these joint effects actually exist, but this would not be a problem if the highly and lowly evaluated countries were similar in these other labour market policies. Regarding PLMPs, in Immervoll et al. (2020) we see that countries like Germany and Sweden are characterized by an intense monitoring of job search but with low levels of sanctioning. Spain and Italy present the opposite situation: an elevated sanctioning with minimal monitoring of job search. Concerning EPL, Germany and Sweden have equal severance pay in temporary and open-ended contracts at least for a worker with 9 months of tenure, while employers in Italy or Spain face higher severance pays in case of open-ended contracts (OECD, 2021)[25]. Lastly, there are also differences in the policies regulating wage-setting. In relation to Germany, it has a collective bargaining rate of 48 %, which is far from the rates found for our countries of interest (Visser, 2024). If we compare the collective bargaining of Italy and Spain versus Sweden, the main difference lies in the ability to enforce the agreements. Following the indicator measuring this phenomenon in Visser (2024), compared with Spain and Italy, Sweden more than doubles the capacity to control collective agreements.

We have shown that there are differences between our countries and the most evaluated ones, and that these differences might impact the effect estimates. Thereby, our dissertation is justified because surveys on ALMPs have not studied

[25] Looking at the EPL index the difference is not there for Spain, but this is a product of the categorization. In Spain, for a worker with 9 months a tenure, an employer pays 50 % of each monthly pay for each month of tenure in case of open-ended contract and 30 % of each monthly pay for each month of tenure in case of temporary contract.

the interrelation between these programs and the rest of labour policies beyond including a categorical variable on the country group in their meta-regressions (Orfao & Malo, 2023; Kluve et al., 2019; Levy Yeyati et al., 2019; Card et al., 2018; Filges et al., 2018; Crépon & Van den Berg, 2016). Sometimes these country or country-year indicators are included, but their estimates are not reported or interpreted (Vooren et al., 2018; Escudero et al., 2018; Card et al., 2010). However, regulations like EPL or CBP have been substantially changing over time (Eichhorst & Marx, 2021; Visser, 2024; Visser, 2016). In light of this situation, ALMP evaluation in Spain and Italy becomes essential.

The second reason is leveraging the importance of regions to fight the unemployment problems suffered in some of them. Spain and Italy have a similar number of regions, with seventeen[26] and twenty respectively. The differences among regions within the countries of the EU-15 have grown since the Great Recession in dimensions like the GDP per capita, except in Germany (Accetturo et al., 2024). In Spain, such divergence is also present if we analyse other indicators like the growth rate in function of the initial level of output (Auciello & Puente, 2023). The case of Italy is even more dramatic, with increasing inequalities from 2006 to 2016 in variables like the employment rates or the number of early school leavers (OECD, 2019b). Nonetheless, citizens of each country think that unemployment is one of three most important issues affecting the *whole* country. This is thought by Italians and by Spaniards both in our period of study (European Commission, 2017) and in the present (European Commission, 2024). The reality is that Spain and Italy have faced large unemployment rates in comparison with the rest of the EU (Eurostat, 2024c).

In this context of regional disparities and common concern about unemployment, an ALMP system mainly managed by regional governments may be seen as an opportunity of mutual learning across administrations. This is similar to the framework promoted by the EU to create benchmarking processes across states aimed at detecting good policies (European Parliament and Council, 2021). Catalonia and Veneto do not have exactly the same economic structure than the less developed regions but offer some innovative programs and infrastructures

[26] There are also two cities, Ceuta and Melilla, with self-government competences. However, they do not have competences to manage ALMPs.

more open to collaborate with researchers. This is an occasion to get some insights on effective policies and try to transfer them to other regions.

The third reason is the low expenditure of Spain and Italy in ALMPs, which requires a more efficient distribution of resources. Although all countries may envisage the maximization of social welfare as their objective, the challenge is more intricated when the resources are scarce and it is necessary a prioritization of users. The low investment in these policies does not seem circumstantial, since Bonoli (2013) has documented how the low expenditure in ALMPs is a long-term feature of Southern European countries. On the one hand, one option might be to demand increases in the budget for these policies. However, Busemeyer (2017) reports how European citizens support more spending on ALMPs like adult training but not at the cost of a decrease in PLMPs. If politicians are reluctant to rise public revenues, public opinion on the funding of labour market polices as a whole complicates an increase in the budget of ALMPs. On the other hand, if we cannot arrive to a high social welfare result by summing more interventions with small effects, we need to detect which are the programs with the largest effects. This is the line pursued by PES like the Catalan, who has a strategic plan to institutionalize policy evaluations as a standard stage in the policy cycle (Ivàlua, 2014). Our intention is to contribute to that effort.

2.2. International cotutelle

This doctoral thesis is related to two subdisciplines: social policy and labour economics. On one side, the literature of social policy has deepened into the design and implementation of labour market policies (Clegg & Durazzi, 2023; Nelson et al., 2022)[27]. Probably due to an academic background more accustomed to dealing with qualitative data (text), researchers in this field have paid more attention to tool or target differences of ALMPs. Clasen et al. (2016) argued that the designs of ALMPs require a more careful attention due to the high number of public administrations and private providers usually involved in their development. On the other side, the literature of labour economics has applied robust methods to

[27] As recognized by Clasen (2013), the social policy literature has a remarkable overlap with the comparative public policy literature. This tradition, more generally called policy sciences, has a greater presence of political scientists. Howlett (2018) is an example of this policy sciences tradition.

estimate more credibly the effects of labour market policies (Card et al., 2018; Cahuc et al., 2014). The better understanding of the properties of randomized experiments has been accompanied with a rise in the publications that try to emulate this research design with observational studies (Imbens & Xu, 2024; Imbens & Rubin, 2015). We argue that joining both traditions is essential to estimate the effect of well-defined interventions (Hernán, 2016). This connection has justified the development of an international cotutelle agreement between the Universitat Autònoma de Barcelona and the Università degli Studi di Padova.

3. Objectives and hypotheses

3.1. Objectives

The doctoral thesis has two general objectives: (1) evaluate the effectiveness of certain tools used in the design of ALMPs, and (2) evaluate the predictive performance of jobseeker profiling models for the design of ALMPs. The first general objective focuses on the means side of policy designs (the tools) and the second one on the goals side (the target groups). These general objectives are further articulated into distinct specific objectives for each chapter.

Chapter II analyses an adult training program with four specific objectives: (1) evaluate the employment effects of the cohort of the program, (2) evaluate the employment effects of the intensity of classroom-based training, (3) evaluate the employment effects of the specificity of adult training, and (4) evaluate the employment effects of the intensity of work-based training. Cohort (timing), intensity and specificity are three design features of adult training. Therefore, this chapter contributes to the general objective 1.

Chapter III studies jobseeker profiling models with two specific objectives: (1) develop a statistical profiling model for jobseekers using only administrative data, and (2) evaluate a rule-based profiling model for jobseekers in comparison to a statistical model. Both specific objectives clearly align with the general objective 2.

Chapter IV assesses a policy mix of job search services with two specific objectives: (1) evaluate the employment effects of endowments with different intensities of job search services, and (2) evaluate the uptake effects of endowments with different intensities of job search services. The intensity is a tool feature both of

job search assistance and of job search intermediation. The uptake deals with the redemption of the amount of a voucher, which has to do with another tool feature. Therefore, these two specific objectives are in line with general objective 1.

3.2. Hypotheses

Each research chapter analyses a different policy, so the hypotheses also vary by chapter.

In Chapter II, we test the following null hypotheses on the effects of a program called FOAP[28]:

- The average treatment effect on the treated on the outcome o after e quarters of participating in FOAP is equal to zero.
    - For $o \in \{employment, stable\ employment\}$ and $e \in \{0,1,2,\dots,12\}$.
- The average treatment effect on the treated on the outcome o during the first six quarters of participating in FOAP at the cohort g is equal to zero.
    - For $o \in \{employment, stable\ employment\}$ and $g \in \{8,9,10,\dots,15\}$.
- The average causal response on the outcome o of increasing from d to j hours the planned duration of FOAP at the e-th quarter is equal to zero.
    - For $o \in \{employment, stable\ employment\}$, $(d,j) \in \{$ 150,250 , 250,500 , 500,750 , $(750,920)\}$ and $e \in \{3,5\}$.
- The average causal response on the outcome o of increasing from d to j the number of occupations targeted as occupational specificity of FOAP at the e-th quarter is equal to zero.
    - For $o \in \{employment, stable\ employment\}$, $d,j \in \{$ 1,2 , 2,4 , 4,6 , 6,8 $\}$ and $e \in \{3,5\}$.
- The average causal response on the outcome o of increasing from d to j percentual points the proportion of internships of FOAP at the e-th quarter is equal to zero.
    - For $o \in \{employment, stable\ employment\}$, $d,j \in \{$ 0,10 , 10,20 , 20,30 , 30,40 $\}$ and $e \in \{3,5\}$.

In Chapter III, we do not test hypotheses using statistical procedures due to the lack of statistical theory for some complex machine learning models or to the high

[28] The target parameters are defined in each chapter.

computational burden of bootstrapping (Athey & Imbens, 2019). In an informal way, using the literature review, we depart from the following hypotheses:

- A statistical profiling model of jobseekers has more sensitivity predicting long-term unemployment than a rule-based profiling model.
- A statistical profiling model of jobseekers needs to incorporate variables on labour market histories in order to achieve a ROC-AUC in the interval of [0.7, 0.8] to predict long-term unemployment.

In Chapter IV, we test the next null hypotheses on the effects of a program abbreviated as AXL[29]:

- The average treatment effect on the outcome o during semester e of being assigned to treatment group r instead of group w is equal to zero.
    - For $o \in \{number\ of\ worked\ days\}$, $e \in \{1,2,3,4\}$ and $r, w \in \{ B, A\ ,\ C, B \}$.
- The quantile τ treatment effect on the outcome o during semester e of being assigned to treatment group r instead of group w is equal to zero.
    - For $\tau \in \{0.25,0.5,0.75\}$, $o \in \{number\ of\ worked\ days\}$, $e \in \{1,2,3,4\}$ and $r, w \in \{ B, A\ ,\ C, B \}$.
- The average treatment effect on the outcome o during the two years since treatment started of being assigned to treatment group r instead of group w is equal to zero.
    - For $o \in \{type\ of\ contract\ of\ longest\ employment\ episode\}$ and $r, w \in \{ B, A\ ,\ C, B \}$.
- The average treatment effect on the outcome o during the implementation period of the voucher of being assigned to treatment group r instead of group w is equal to zero.
    - For $o \in \{hours\ of\ job\ search\ assistance, hours\ of\ adult\ training\}$ and $r, w \in \{ B, A\ ,\ C, B \}$.

4. Methodology

In this section, we present the key methodological aspects used during the research process. First, we are going to discuss three methodological decisions that are present along all the dissertation: our causal inference framework for internal

[29] This is estimated on a specific subpopulation. See the details in Chapter IV.

validity, our causal inference framework for external validity, and our framework of statistical inference. Second, we briefly describe the analytical strategies developed in each chapter to answer our research questions.

4.1. Common to all chapters

Firstly, to guarantee the internal validity of effect estimates, we depart from the potential-outcomes framework of causal inference. The theoretical literature on causal inference may be currently grouped in two schools: the potential-outcomes (PO) and the structural causal model (SCM) approaches. This division is sometimes summarized with the seminal works of Rubin (1974) and Pearl (2009). Both traditions are not completely separated, since recent works on econometrics have included insights from the SCM school (Abadie & Cattaneo, 2018; Frölich & Sperlich, 2019; Cunningham, 2021; Imbens, 2024). Each literature puts different emphasis on certain topics, e.g. the PO school has paid much more attention to the overlap assumption than the SCM researchers (Imbens, 2020), while the SCM literature has invested more effort in thinking on the specific set of covariates necessary for unconfoundedness to hold. Moreover, understanding is sometimes complicated by using different words to talk about similar or almost equal concepts (Pearl, 2009), e.g. PO researchers talk about overlap or common support, while SCM authors call it positivity[30]. Hernán & Robins (2023) have indicated the several common points of both approaches and have tried to join them, but the adoption of their proposal is still low in the social sciences[31]. In this doctoral thesis, to follow the current trend in social sciences, we use the potential outcomes framework with the refreshed version of its notation that is present in works like

[30] A more nuanced terminological difference has to do with the concept of unconfoundedness. The PO school considers that unconfoundedness is a property of certain assignment mechanisms (Imbens & Rubin, 2015), whilst the SCM school argues that unconfoundedness is a property of certain effect estimates (Pearl, 2009).

[31] This literature is placed between the PO and the SCM approaches and has a great presence of epidemiologists. These researchers have faced domain-specific problems like a proper definition of variables whose effects wants to be measured, e.g. obesity. This drove them to state that there is an assumption typically forgotten in causal inference that they have called the consistency assumption (Hernán & Robins, 2023). This assumption may be split in two statements: (1) well-defined interventions and (2) realizations of these interventions in our dataset. This pushes researchers to describe the theoretical intervention in a detailed way and then to contrast if this is the intervention they are analysing in their study. As we will see during the doctoral thesis, this assumption is intimately linked to the design approach that runs through all our work.

Frölich & Sperlich (2019). The unique difference is that this updated notation sometimes includes causal graphs or arrows like $D \rightarrow Y$ to denote the effect of D on Y.

Secondly, to interpret the results of other studies we also depart from a causal inference framework adapted to the external validity problem. In this dissertation, the references to the effects estimated in the literature for similar programs must not be considered a mental experiment on the effect that we would have found if we would have implemented such intervention in our population of interest. As explained by Pearl & Bareinboim (2014), this exercise requires a more nuanced estimation process specifying the assignment mechanisms in each case and if interaction effects are assumed. If the assignment mechanisms or the interaction assumptions of our intervention and a third intervention do not coincide, we need post-hoc statistical adjustments. For instance, it is conventional to assume that *age* is an effect modifier of the impact of ALMPs, and we know that in the assignment mechanism of some programs *age* influences the treatment probability (Katz et al., 2021). Thereby, to transport the results of a study like Katz et al. (2021) to our context, we would need to adjust for age differences. An additional difficulty is that it is complicated to know if the two interventions are identical, since the design and the implementation are usually poorly described.

Thirdly, the statistical inference carried out in all analyses is with respect to a superpopulation. Berry et al. (2014) describes that researchers may pursue statistical inference from a population mode of inference or from a permutation mode of inference. On one hand, experimental researchers classically relied on a mode of inference that tries to account for the random allocation of treatments across individuals, considering the specific individuals as something fixed[32]. On the other hand, observational studies have usually used sampling-based inference. This mode of inference tries to estimate the uncertainty coming from the random selection of individuals (samples) of a certain population, assuming that the specific distribution of treatment is something fixed. Due to this focus on different sources of uncertainty, in the permutation mode of inference the sample is the whole population of interest.

[32] By fix we mean that the uncertainty linked to its distribution is not considered in the inference techniques.

In our case, although we use the universe of *all* participants in a certain edition of program, we may still understand this population as a sample of a superpopulation. We are interested in inferring the results of our research to a population broader than the universe of people treated in a certain call. For the evaluation in Chapter II, the target population is the set of all persons willing to participate in ALMPs in Catalonia (Spain). Equivalently, the target population in Chapter IV is made up by those willing to participate in ALMPs in Veneto (Italy). This superpopulation is more modest than a superpopulation of all jobseekers in a certain region. Having a sample that resembles the target population in certain features of its multivariate distribution is necessary to have a consistent estimation. The problem is that, since the true values of the target population are unknown, claiming that a sample resembles a population implicitly involves assuming a certain mechanism on the missing individuals (Mohan & Pearl, 2021).

We argue that considering the "potential participants" as the target population of the inferences is a less restrictive assumption. The assumption is that the decision to participate ($W_i = 1$) is affected by some variables related to the potential outcomes ($\boldsymbol{G}_i$) and some other variables unrelated to the potential outcomes once accounted for the last ($\boldsymbol{\epsilon}_i$). Formally, the assumption is: $\boldsymbol{\epsilon}_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}|\boldsymbol{G}_i$. Our sample does not include data on some jobseekers that might participate but did not because $\boldsymbol{\epsilon}_i$ is a vector of factors that occur by chance. For example, $\boldsymbol{\epsilon}_i$ might have a variable collecting if the individual saw a poster on the PES, considering that those posters were placed randomly around the city. Thereby, we conceptualize the inclusion of members of the mentioned population in our sample as a missing at random process.

4.2. Specific to each chapter

Chapter II is aimed at estimating the effects of an adult training program called «Formación de Ofertas en Áreas Prioritarias». To identify the causal estimands, it uses a conditional difference-in-differences strategy. Since the assignment mechanism of the program is not clear, we also employ a conditional lagged dependent variable strategy to bracket the effect estimates. We include placebo outcome analyses as plausibility checks on the identification assumptions. The

additional results come from robustness analyses that varied another identification assumption, the so-called parameter of anticipation.

Chapter III is dedicated to evaluating a rule-based profiling model for jobseekers called «Q» and to developing a statistical alternative. The objective of these models is prediction, so we estimate discrimination and calibration statistics with four modelling strategies from less to more flexible: unpenalized logistic regression, penalized logistic regression, a random forest model, and a gradient boosting model. As additional findings, we incorporate the results from two stress tests: a shifted performance evaluation and a stratified performance evaluation.

Chapter IV examines the effects of a job search services program called «Assegno per il Lavoro». To identify the causal estimands, it employs a sharp regression discontinuity strategy. We study the plausibility of the identification assumptions with analyses of score manipulation, placebo outcomes, and covariate balance. The robustness checks attend if variations in the bandwidth, the kernel, the polynomial degree, or the population of interest generate large changes in the conclusions of the research.

Chapter II.

Employment effects of adult training in Catalonia: heterogeneity by design[33]

1. Introduction

Although adult training is the oldest active labour market policy, we still do not fully understand why its employment effects are so varied (Bonoli, 2013; Levy Yeyati, 2019). Meta-analyses of evaluations report a standard deviation that almost doubles the mean estimate (Card et al., 2018), and, if we differentiate by time horizon, we find standard deviations that even multiply it by six (Vooren et al., 2019)[34]. Previous research has discussed several candidate variables to explain this variability (Kemper et al., 2022). They can be broadly grouped in factors related to the characteristics of the intervention, the features of the participants, and the evaluation decisions. Whereas the heterogeneity analysis focused on features of the population of participants is more common (e.g. differences by sex or level of studies), the remaining factors have received limited attention in impact evaluations.

In this chapter, we evaluate an adult training program implemented in Catalonia (Spain) by focusing on different sources of heterogeneous effects. On one side, we analyse *time* dimensions, either the moment the outcome is observed or the moment the treatment begins. On the other side, we analyse the *tools* of the intervention by looking at the design features of the different components. Our research questions study how estimates of employment effects vary according to these factors, looking at both the probability of employment and the probability of stable employment. The period of outcome measurement is a decision of the evaluator, but we have evidence for other countries on its importance due to the

[33] This chapter is based on a working paper titled "Heterogeneity by design? Heterogenous treatment effects of adult training in Catalonia", which is authored by me.

[34] Using the dataset of estimates of Card et al. (2018) on employment effects of adult training, the mean estimate is of 4.71 p.p. and the standard deviation is of 8.59 p.p. Vooren et al. (2019) reports that 12 months after program start, the mean estimate is of -0.02 p.p. and the standard deviation is 12.3 p.p.

dynamic effects of adult training. The timing of a treatment (the cohort) might be an effect modifier due to the high seasonality of labour markets in Southern Europe, which could make training providers to schedule their treatments trying to catch this premium. Lastly, we have analysed program tools to measure for each version of the treatment its planned duration, its occupational specificity, and its proportion of internships. When both elements are considered simultaneously, the first tool can be interpreted as the intensity of classroom-based learning, while the latter represents the intensity of work-based learning. Design variability has been discussed at length in the literature[35] and there are theoretical reasons to believe that different dosages of these components might trigger different changes in the outcomes.

Methodologically, we adopt an agnostic position on the identification assumptions regarding the model of untreated potential outcomes. Although we base our main analysis on a conditional difference-in-differences strategy, we compare the results obtained with those produced with a conditional lagged dependent variable strategy. Ding & Li (2019) show that this combination allows us to bound the effect estimate of an intervention with less stringent assumptions. From these identification assumptions, we estimate models operationalizing treatment participation as either a binary variable or three continuous variables collecting the dosage of each component. This last configuration understands our program as a complex treatment defined like a factorial design (Wu & Hamada, 2021) or a compound treatment (Hernán & VanderWeele, 2011).

The main results of the three research questions were the following. Concerning the dynamics of the effects, the estimates are monotonically increasing over time but stabilize around 26 p.p. for the employment probability and 18 p.p. for the stable employment probability after two years. The conditional lagged dependent variable analyses maintain the statistically significant positive effect and smooth

[35] Back in 2006, the OECD Employment Outlook summarized the ALMP micro-evaluations concluding that: "one of the clearest lessons is that the returns to different programmes vary widely and that apparently similar programmes can yield very different outcomes, implying that the details of programme design are key" (OECD, 2006, p. 72). After them, numerous researchers have highlighted the great diversity of program designs labelled as adult training (Crépon & Van den Berg, 2016; Levy Yeyati et al., 2019; Kemper et al., 2022). A very recent review still indicates the same concern: "The plentiful variation in the design of ALMP programmes has not received sufficient attention (…) A general takeaway from this literature is that we need to take such design variations seriously" (Cronert, 2023, p. 347).

the presence of dynamic effects, but estimate that at that moment the impact would be of 5 p.p. and 3 p.p. respectively. Regarding the cohort effects, we do not have conclusive evidence to assert clear differences in effects among cohorts. Confidence intervals usually overlap for both outcomes, but we do find some correlation between estimates of cohort parameters and economic growth. The results with the alternative identification strategy also indicate small differences between the point estimates of cohorts for both outcomes, which again calls into question the importance of cohort effects. With respect to the impact of the different tools, our results challenge previous studies on the effect of planned duration on employment by finding a convex instead of a concave relationship at least after one year and a half. Our interpretation is that previous research might be confounding the effect of duration with the effect of the occupational specificity or the proportion of internships. Although we detect positive effects of participating in a highly specific treatment, the largest impacts are found for increases in the proportion of internships, especially comparing courses with 0 % to 10 % of internships.

The contributions of our research are both substantive and methodological. Firstly, for the time analysis, we advance the literature by using less stringent identification assumptions to evaluate the impact of adult training in Southern Europe. As shown in Card et al. (2018) and Vooren et al. (2019), there is scarcity of ALMP evaluations in these countries. Since they present configurations of labour market policies that do not generally resemble countries more studied like Germany, we argue that they deserve more attention. Moreover, in these setups previous research had been only based on a conditional lagged dependent variable strategy (Cueto et al., 2010; Clemente et al., 2014). Leveraging the new methods of Callaway & Sant'Anna (2021) for staggered designs, we are also able to estimate cohort effects properly. Secondly, for the tool analysis, our study is the first to analyse the effect of three design components by controlling the simultaneous implementation of these measures. Studying each component separately bypasses that the presence of one component may be correlated with the presence of another, e.g. longer treatments having greater proportion of internships. This one-at-a-time analysis inevitably aliases (confounds) the effect of one component with the effect of the other. Moreover, following previous results, we have also relaxed the linearity assumption on the dose-response function by specifying quadratic terms.

The following section frames our research in the literature of adult training and describes the design of our program. Section 3 presents the data and the methods used to analyse them. Section 4 reports our findings by source of heterogeneity and offers some robustness analyses to assess if our results change according to different assumptions. To finish, the last section includes the conclusions with possible lines of future research.

2. Adult training

2.1. Literature review

The first question we address regards the effect of adult training programs on the probability of finding employment. Looking at experimental data, Card et al. (2018) reports that the median estimate is an average increase of 2 p.p. of the employment probability. However, as we mentioned in the introduction, there is a remarkable heterogeneity, since the range of estimates exceeds 20 p.p. at the short and medium terms[36]. Using a much bigger sample of experiments, the meta-analysis by Levy Yeyati et al. (2019) finds a median impact of 6.7 p.p., joint with an important variability as well. In both reviews, almost all interventions started in this century, but more than half of the sample comes from non-OECD countries. The second question we study deals with the impact of adult training on the probability of finding a stable employment, which has been much less studied[37]. Dengler (2019) found for Germany that adult training increased the probability of stable employment by 2.3-4 p.p. in an interval of 33 months since treatment started. Wood et al. (2024) reports an effect of around 5 p.p. after two years for classroom-based training in Belgium.

Some articles suggest that the variability of effect estimates might be influenced by the moment of measurement of the outcome (i.e. dynamics of effects) or the starting date of the treatment (cohort effects). Regarding dynamics, meta-analyses have found that employment effects of training tend to be higher at the very long run (Vooren et al., 2019b), but the increases do not seem large in

[36] Based on Card et al. (2018), we define the effect horizons in the following way: very short run = [0, 6 months], short run = [7, 12 months], medium run = [13-24 months], and long run = [25, ∞ months). The original categorization in Card et al. (2018) collapses the two first categories.

[37] Unlike the outcome of employment, to our knowledge, there are no systematic reviews on the effect of ALMP on job quality outcomes related to the type of contract.

absolute terms (see Kemper et al., 2022, who found non statistically significant rises for youth). We will provide a more detailed analysis of these dynamics with a quarter-based dataset.

Concerning the moment of implementation, the main finding of Card et al. (2018) and Kluve (2010) was a higher effect of ALMP in recessions than in expansions. Corrales-Herrero & Rodríguez-Prado (2018) confirms this for Spain assuming selection on observables. However, Casanova et al. (2021) found using difference-in-differences (DID) that the employment effect of similar training programs went from 4 p.p. to 10 p.p. when going from 2011 to 2016, a period dominated by economic growth. The business cycle has been the classical explanation for such differences. Our detailed dataset allows us to focus on seasonality, which has been rarely studied in the ALMP literature. Lechner & Wunsch (2009) is an exception, getting results indicating a small and statistically non-significant relationship between effect size and quarterly GDP growth. We will relax their selection on observables assumption by using an alternative identification strategy.

Other authors have tried to understand the heterogeneity of results looking at the tools included in the interventions. Our program is very similar to the so-called sector-focused training programs reviewed in Katz et al. (2022). The designers of these interventions only fund courses that target jobs predicted with high local labour demand at the short and long term. They reported that in the medium run this type of program caused around a 5 p.p. increase in the employment probability. Besides the careful selection of target jobs, we have studied three design features of the different tools included in the program.

The first design variable is treatment duration. The international literature presents mixed evidence about this topic: Vooren et al. (2019b) finds that in general there is no relationship with the employment probability, although Levy Yeyati et al. (2019) reports a positive correlation and Kemper et al. (2022) also estimates a positive association for younger people. In Spain, Cueto et al. (2010) detected that the effect was an increasing function of the planned duration of training. The results of Clemente et al. (2014) for the medium run qualified this conclusion by finding, in general, minor changes between intermediate and long courses[38]. While these findings use data on treatments during the expansion of the

[38] Note that Cueto et al. (2010) and Clemente et al. (2014) categorized in a different way the planned duration of a course. See Appendix A for more details.

business cycle, Blázquez et al. (2019) also found a nonmonotone shape of the relationship for years of recession. There are two additional studies that considered actual treatment duration as a continuous factor and used weeks as scale. For USA, Flores et al. (2012) estimated that the effect of treatment hours on the wage was increasing four years after treatment started. For South Korea, Choe et al. (2014) only found a small effect of 1.72 p.p. when comparing very long with very short courses. The rest of comparisons triggered not statistically significant results for completers. Concerning the outcome of stable employment, Crépon et al. (2012) estimated for France that longer training treatments triggered longer employment episodes. This is consistent with Osikominu (2013), who also found that higher duration caused higher employment stability.

The second design variable is the occupational specificity of each treatment. Occupational specificity is understood as the degree of linkage between one training credential and a set of occupations (DiPrete et al., 2017). When the treatment is connected to a few occupations in the labour market, its linkage is strong (highly specific); whereas when related to many occupations, its linkage is weak (lowly specific). Since policymakers typically lack this information, we distinguish between actual and perceived specificity[39]. A recent strand of labour market research has focused on the effect of actual specificity on school-to-work transitions (Hanushek et al., 2017). Bol et al. (2019) found with granular data that the actual occupational specificity of training credentials was positively associated with the employment probability in France, United States, and Germany. Recent research supports the idea that these positive effects decrease in the very long run (Rözer & Bol, 2019). DiPrete & Chae (2023) have extended this research line to more countries in Europe and have also found that more linkage is related to more employment probability. They report that this effect is substantially lower for older women. Wood et al. (2024) also found that, in contrast to low specific treatments, high specific training involved positive effects after two years in our two outcomes of interest. Our aim is to expand this literature by using a less strict identification strategy with data on perceived

[39] Establishing a parallelism with the overeducation literature (Hartog, 2000), the linkage might be measured both through experts' analysis (trying to predict or describe a pattern) and through realized matches. The first would be the perceived linkage, and the second the actual one. We use a novel dataset that gathers the perceived specificity of each treatment as indicated in its regulation, that is, as defined by experts.

specificity and by analysing whether there are particularities for adult instead of initial training.

The third design variable is the proportion of internships. Following Sanz de Miguel et al. (2021), we use internships or traineeships to talk about work-based training *without* a labour contract and we reserve apprenticeships for work-based training with a labour contract[40]. This component has been studied in relation to programs for young people, although impact evaluations are scarce (Lodovici et al., 2013). The studies with the strongest identification strategies for this age group found that they seem to impact positively the employment probability (Margaryan et al., 2022; Baert et al., 2021). In Spain, only apprenticeships for young workers have been evaluated. Specifically, Bentolila et al. (2023) found positive effects in dual vocational education programs and Mato-Díaz & Escudero-Castillo (2023) also estimated a positive effect in the frame of an ALMP accompanied by employment subsidies. Wood et al. (2024) studied the impact of internships and estimated that they were much more effective than apprenticeships both on the employment probability and on the stable employment probability.

2.2. Design of the program

«Formación de Oferta en Áreas Prioritarias»[41] (hereinafter, FOAP) is a program funded by the public employment office of Catalonia. It is the main adult training program of Catalonia (Spain). In this chapter we focus on the calls executed from November 2016 to October 2018 (SOC, 2016; 2017). The general objectives of the program were two: improving the level of education and increasing the probability of employment (Government of Catalonia, 2016a).

The main tool of the intervention was classroom training. In some cases, this component was complemented with work-based training through internships. This last element is mandatory for those treatments that grant a certificate recognized

[40] The first tool might induce rises in employment probability due to: (a) increases in productivity (from the output side) or (b) a better screening of the candidate for employment at the trainer firm (González Veiga et al., 2008). The second tool might also explain rises in employment probability with decreases in labour costs (input side of productivity), since they are usually accompanied by subsidies (Cueto & López, 2020).

[41] We might translate it to English as "supply-side training in priority areas".

in the whole state, the so-called professional certificate (*certificado de profesionalidad* in Spanish). As we will see, the intensity (duration) of both components varies by course. The distinctiveness of FOAP is that designers carefully select the training content according to labour and product market trends. This follows a similar logic to the sector-focused training programs in USA (Katz et al., 2022).

The target population of our program are individuals who are unemployed. The focus on this collective is ensured by partially limiting the participation of people already employed up to the 30 % of the treatment group. Furthermore, the design establishes that individuals with certain features should be prioritised. Some of these characteristics are having a disability, a low educational level, or a very long unemployment episode.

Concerning the implementers of the intervention, this program establishes public funding but allows both private and public provision of the services. Private organizations and local public administrations may apply for funding if they accomplish certain technical requirements established for each course. The grant is based on a score derived from the territorial priority of the specific course, past implementations, the human resources of the organization, and the engagement with the jobs related to the course.

It is also necessary to consider the rules governing FOAP. The different courses may start and end at different dates of the year, being this a decision made by the implementer. This is grounded in the countercyclical character of ALMP, offering a quick response to labour market shocks against the rigid calendars of the educational system. Although the number of total hours of a course is set up by the central government, the calendarization of lessons is set by each provider. Fixing the calendarization, if training implied a break of the job search (the lock-in effect), different starting dates might provoke different employment probabilities due to the seasonality of some occupations. Fixing the starting date, it is also reasonable to think that courses with low weekly hours of training (so longer total duration in months), might hardly stop the job search. This planned duration in months or weeks is something that unfortunately we do not observe.

3. Methods

3.1. Data

To answer our research questions, we have used a rich file of administrative data. This dataset merges information from databases on the intermediation claims (SICAS), on the features of the treatments (GIA), on the employment contracts (CONTRAT@), on the Social Security affiliations (AFI), and on the planned duration of treatments in terms of number of hours (Dades Obertes Catalunya, 2023). Except the last one, all of them are confidential and have been on loan from the public employment service of Catalonia. An additional dataset has been produced through scraping of the website of the public employment service of Spain (SEPE). It links each treatment to its target occupations as established in the regulation of each credential. A detailed description on all the variables involved in our study is available at Appendix A, joint with a table collecting the summary statistics of these covariates in our sample. Here we offer the main characteristics of our variables.

We considered two outcome variables at different time horizons: having an employment and having an open-ended contract. Having an employment ($Y_1 \in \{Yes, No\}$) collects if the individual worked the last day of quarter t or if signed an employment contract at any day of such quarter. Having an open-ended contract ($Y_2 \in \{Yes, No\}$) is usually considered a dimension of job quality (Muñoz de Bustillo et al., 2011) and records if the person had an open-ended contract at the end of the quarter or if signed at least one of them during the quarter. This outcome measures the planned stability of the employment, which is different from measures of actual stability of the employment like in Dengler (2019). Treatments occur between 2016Q4 and 2018Q4, but we have information on the outcomes for the temporal window 2015Q1 – 2019Q4.

The treatment variable is defined in two ways. For the analyses of time-based heterogeneity, it is simply a binary variable indicating the participation in a FOAP course. For the analyses of tool-based heterogeneity, we distinguish three treatment variables because we look at three dimensions of the intervention: planned duration, occupational specificity, and proportion of internships. Firstly, the planned duration of the treatment has been operationalized through the number of hours declared in the regulation of each treatment. Secondly, the concept of occupational specificity has been measured by counting the number of

distinct occupations (classified with CNO-11 at 4 digits) linked to each treatment in its regulation[42]. Thirdly, the proportion of internships is established in the design of the programs. For those trainings related to a professional certificate, we collected the planned duration of the internships. The trainings that do not grant a professional certificate are a minority, but for these cases we had to take the actual proportion of internships since they do not fix a minimum proportion by design. Additionally, we also adjusted for a fourth component: training linked to a closed occupation. Following the results of Bol & Weeden (2015) on occupational closure, we also considered important to remove those credentials targeted towards at least one occupation with regulated access. We identified these credentials with the dataset offered by SOC (2023). We focus on the first three factors because we think that they are the ones that have more policy interest for a broader audience[43].

Our data also include control variables. Since we deal with observational data, we had to perform adjustments on differences in the distribution of confounding variables between treatment groups. We have information on the sociodemographic features of the individuals and on their labour market histories (including the histories of unemployment benefits). We will go more into detail when discussing the identification strategy and Appendix A offers exhaustive information on them. Before regression adjustment, we also applied four restrictions on our sample to delimit our population of interest: (1) to remove those treatments that were started at previous editions of FOAP[44], (2) to remove those treatments that were begun while being employed, (3) to remove those treatments that were started during an employment subsidy treatment, and (4) to remove those treatments exclusively devoted to people with disability. These restrictions make easier the identification of the counterfactual through a less difficult specification of the causal model.

---

[42] One strand of the literature has proxied specificity by measuring how concentrated are the empirical distributions of occupations for each treatment, i.e. through realized matches. They could estimate it because they dealt with a relatively short list of possible treatments (24 fields in Bol et al., 2019; and DiPrete et al., 2017) or because they used very granular data (Forster and Bol, 2018). In Spain, there are not detailed microdata on field of study joint with occupation. Since the literature emphasizes differences between countries, we opted for an operationalization based on the linked occupations defined by policy designers.

[43] Closing occupations does not seem a reasonable intervention just to foster employment, so we exclusively focus on the strata of treatments not targeted to closed occupations.

[44] This is possible due to the modular design of the training certificates.

3.2. Analysis

3.2.1. Target parameters

The first analysis focuses on time heterogeneity and uses the adaptation of the potential-outcomes notation to the dynamic and staggered setup proposed by Callaway and Sant'Anna (2021). Treatment is formulated as a binary variable called D_{it} to register if individual i was treated in time t. The potential outcome of individual $I = i$ in period $T = t$ is denoted as $Y_{it}(g)$ if she was first treated in period $G = g$ and as $Y_{it}(0)$ if she was never treated in our temporal window. Both the time and the group (cohort) variables are defined like indexes as $T \in \{1, ..., t_T\}$ and $G \in \{0, ..., g_G\}$, denoting with $G = 0$ those individuals never treated in our temporal window. Later we will need $G_{ig} = \mathbf{1}\{G_i = g\}$, which simply indicates if individual i belongs to cohort g. To reduce the notational burden, we remove the individual subscript.

From this framework, we define two target parameters that address different possible dimensions of heterogeneity in the effects: the elapsed duration since treatment started (dynamics) and the cohort effects provoked by a staggered design. Both estimands have as building blocks the average treatment effect on the treated since time g at time t, $ATT(g,t)$, proposed by Callaway and Sant'Anna (2021). Briefly, this parameter is defined as:

$$ATT(g,t) = \mathbb{E}[Y_t \ g \ | G = g] - \mathbb{E}[Y_t \ 0 \ | G = g]$$

It works as an auxiliary parameter since it is not the focus of our research but needs to be estimated. From them, we construct the average treatment effect on the treated after $e + 1$ periods (θ_e) and the average treatment effect on the treated in cohort g after $e + 1$ periods ($\theta_{g,e}$)[45]. First, define $\mathcal{G} = \mathsf{supp}(G)\backslash\{0\}$ as the set of possible G values excluding the never-treated group. Now, we can define them formally as follows:

[45] Note that this last parameter is slightly different from the one presented in Callaway and Sant'Anna (2021) to check how ATT's vary across groups. In the presence of effect dynamics, to isolate the cohort effect we need to equate the number of periods introduced in the average for each group. Remind that we talk about the effects after $e + 1$ periods because the first quarter since treatment started is $e = 0$.

$$\theta_e = \sum_{g \in \mathcal{G}} \mathbf{1}\{g + e \leq t_T\} \Pr[G = g | G + e \leq t_T] \, ATT(g, g + e)$$

$$\theta_{g,e} = \frac{1}{e+1} \sum_{t=g}^{g+e} ATT(g, t)$$

The second analysis of FOAP defines it as an intervention with different components. The key point is that the components of interest have infinite possible levels, so we think them as continuous treatments. Therefore, we reserve the treatment variable $D_{it} \in \{1,0\}$ to flag those individuals that participated in any version of FOAP and define a component variable to gather the dose of component k received by individual i at time t, so $\widetilde{D}_{itk} = d \in \mathbb{R}$ s. t. $d \geq 0$. For notational simplicity, we remove the unit and time subindexes. Note that this creates an infinite set of potential outcomes $Y(\widetilde{D}_k = d)$. As we will motivate later, we will fix the elapsed duration since treatment started and we will estimate these quantities on the population of individuals ever treated in FOAP ($D = 1$). Inspired by Bia et al. (2014) and by Callaway et al. (2024), we are interested in two target parameters: the average dose-response function of component k in period e ($ADRF_e^{\,k}$) and the average causal response of receiving d units instead of l units of component k in period e ($ACR^{(k)}_{dj,e}$). Formally, we define them as:

$$ADRF^{(k)} = \mathbb{E}\left[Y\big(\widetilde{D}_k = d\big)_e\right]$$

$$ACR^{(k)}_{dj,e} = \mathbb{E}\left[Y\big(\widetilde{D}_k = d\big)_e - Y\big(\widetilde{D}_k = j\big)_e\right]$$

The $ADRF^{(k)}$ is simply the function of treated potential outcomes for different versions of this treatment. The $ACR^{(k)}_{dj}$ is a contrast parameter, since it compares two potential outcomes.

3.2.2. Identification

The target parameters contain counterfactual potential outcomes, so we need identification assumptions to estimate them with observable data (Imbens & Rubin, 2015). The credibility of the identification assumptions is usually judged in relation to the properties of randomized experiments, as the latter have these conditions due to their own design (Rubin, 2007). Let us review how we dealt with these requirements in the two analyses of our study.

The first identification assumption is the unconfoundedness of the assignment mechanism. Qualitatively, we depart from a model of the treated potential outcome in which the treatment variable is a function of observables and unobservables. The particularity of this model is the specification of a concrete set of confounding variables that are observed in our dataset (vectors $\boldsymbol{S}_i$ and $\boldsymbol{L}_i$) plus a set of confounding unobserved variables ($\boldsymbol{U}_{it}$) that we think we can proxy. $\boldsymbol{S}_i$ collects a list of sociodemographic variables, $\boldsymbol{L}_i$ a set of labour market variables, and $\boldsymbol{U}_{it}$ a list of latent variables collecting a linear combination of time-invariant unit-specific covariates ($\boldsymbol{U}_i$) and unit-invariant time-specific covariates ($\boldsymbol{U}_t$). A detailed explanation of these confounders is available at Appendix A. For notational simplicity, we pool all the observed control variables in matrix $\boldsymbol{X}_i$. Following Caetano and Callaway (2023), we condition on time-invariant covariates and avoid time-varying variables that would impose stronger functional form assumptions. We will test the sensitivity of the results with a lagged dependent variable strategy by using the bracketing relationship found by Ding & Li (2019).

On the one hand, for the analysis of time-based heterogeneity, the unobservable quantities are approximated through a conditional difference-in-differences (CDID) based on not-yet-treated individuals (Callaway & Sant'Anna, 2021). The DID part assumes that the amount of confounding bias between groups is stable over time (Fröhlich & Sperlich, 2019). The conditional version of this assumption allows confounders to have a time-variant effect on the treatment or the outcome variable (Zeldow & Hatfield, 2021)[46]. To foster clarity, we are going to split the exposition on the unconfoundedness of the treatment assignment in one expression for the sub-assumption on the base period (Assumption 1) and other expression for the full assumption (Assumption 2a)[47].

[46] As shown in Zeldow & Hatfield (2021), adjusting for time-invariant covariates in a DID study with only one regression equation only makes sense if we added interactions between such covariates and the time indicators. In a staggered DID study à la Callaway & Sant'Anna (2021), we are estimating many regression equations (one for each $ATT(g,t)$), so we are allowing the "effect" of such covariate to vary at each group-time.

[47] As is common in the literature, we also assume the stable unit treatment value (SUTVA).

Assumption 1. There is a known $\delta \geq 0$ such that

$\mathbb{E}\big[Y(g)_t | G_g = 1, X\big] = \mathbb{E}\big[Y(0)_t | G_g = 1, X\big]$ a.s. for all $g \in \mathcal{G}$ and $t \in \{1, \dots, t_T\}$ s. t. $t < g - \delta$.

For Assumption 1, it is key to interpret δ either as the periods of treatment endogeneity, being or not a confounder, or as a proper anticipation effect, whose effect should be accounted when forming aggregated parameters. This assumption is especially important when studying ALMPs, since their eligibility rules usually require that treatment groups be in a bad labour market situation before treatment, generating a pre-treatment difference in the outcome variable called Ashenfelter's dip (Heckman et al., 1999)[48]. In line with the literature, our starting point will be to consider the Ashenfelter's dip as an expression of endogeneity with transitory differences (Malani & Reif, 2015). Thereby, we will assume that the stable bias is best approximated with the last period without group differences in the outcome $(t - \delta - 1)$[49]. We will assess how sensitive are our results by assuming that pre-treatment differences are permanent instead of transitory.

[48] The Ashenfelter's dip was first reported in Ashenfelter (1975). Thinking on endogeneity, Ashenfelter (1978) originally framed the discussion in terms of transitory versus persistent differences in the untreated potential outcome. Card & Sullivan (1988) clarified a bit more the structural interpretation of such pre-treatment differences by talking about the lagged outcome as a proxy (descendant) of an economic shock. Beyond ALMPs, the question on which variables should be adjusted to remove confounding was clarified much later with the publication in the year 2000 of *Causality* (Pearl, 2009). Thereby, the question is if the variable proxied (shock or another unobserved factor) acted as a confounder ($D_t \leftarrow Y_{t-1} \rightarrow Y_{t+1}$) or not ($D_t \leftarrow Y_{t-1}$). If the effect of the shock was transitory, i.e. if there was no effect of the variable proxied by Y_{t-1} on Y_{t+1}, there wouldn't be confounding ($D \leftarrow Y_{t-1}$). If the effect of the shock was persistent, there would be a backdoor path if it also affected treatment participation, i.e. there would be confounding if Y_{t-1} it is not adjusted ($D \leftarrow Y_{t-1} \rightarrow Y_t$). Thinking on the pre-treatment differences as anticipation, we would need to flip the direction of the arrow: $D \rightarrow Y_{t-1}$. In any case, when the analyst departs from a conditional two-way fixed effects model of the untreated potential outcome, it is implicitly ruling out that confounding came from time-variant differences in the unobserved confounder. Therefore, the question reduces to: which pre-treatment difference best proxy the time-invariant unit differences in post-treatment periods?

[49] Appendix B includes more detail on the selection of the anticipation parameter. Beyond the stable bias approximation, DID regressions require much stronger assumptions to attain a consistent estimation if we include lagged outcomes as control variables (Angrist & Pischke, 2009). This is the usual reason advocated to reject its inclusion in the specification.

Assumption 2a. For each $g \in \mathcal{G}$ and each $t \in \{2, \dots, t_T\}$ such that $t \geq g - \delta$,

$$\begin{aligned}\mathbb{E}\big[Y(0)_{t-\delta-1} | G_g = 1, \boldsymbol{X}\big] - \mathbb{E}\big[Y(0)_{t-\delta-1} | i \in \mathcal{C}_{gt}, \boldsymbol{X}\big] \\ = \mathbb{E}\big[Y(0)_t | G_g = 1, \boldsymbol{X}\big] - \mathbb{E}\big[Y(0)_t | i \in \mathcal{C}_{gt}, \boldsymbol{X}\big]\end{aligned}$$

Assumption 2a establishes bias stability (common trends) and delimits two requirements to enter in the control group. First, being part of the not-yet-treated cohorts. Thereby, having fixed g, $\mathcal{N}_g = \mathsf{supp}(G) \backslash \{1, \dots, g\}$ is a set that collects those treated in $G > g$ or in $G = 0$. Second, not having been already affected by treatment at the period of interest, so being in $\mathcal{C}_{gt} = \mathcal{N} \backslash \{t \geq G - \delta\}$. This last set collects those that passed the first test, but having also removed those whose period of beginning of effects was equal or lower to the period of interest. In a nutshell, those that have the treatment switched on. Notice that for notation simplicity, we joined all vectors of confounding variables in matrix $\boldsymbol{X}$. Since FOAP can actually take different forms, we can think D as a binarization of a multidimensional variable, so Assumption 2a may be understood as the aggregate parallel trends assumption formulated in Callaway et al. (2024).

On the other hand, for the analysis of tool-based heterogeneity, the program tools are considered different continuous treatments[50], so identification requires more stringent assumptions. We maintain Assumption 1, but Assumption 2a needs to be reviewed. Callaway et al. (2024) show that in the case of a DID strategy, we need what they call strong parallel trends assumption. Our setup has two features that deviates from its prototypical assumption. On one side, we also assume that we need other observed confounders for identification, conforming the assumption of strong parallel trends conditional-on-covariates. In the spirit of Yu & Ding (2023), to isolate the effect of one component, we additionally need to adjust for the levels of the other design features. We may think in a matrix $\boldsymbol{K}$ collecting the values of the non-focal components (design features), but for notational simplicity we pool all controls in $\boldsymbol{X}$. On the other side, since we now distinguish a myriad of potential treatments, we consider that identification is more credible if we recognize that there are unobserved confounders for the effects of interest. Following Bell et al. (1995), we argue that only considering the population of

[50] To be more precise, we could talk about a multi-valued discrete treatment, since we have some treatment values with a remarkable probability mass. Following Callaway et al. (2024), the main identification arguments are not affected by this subtle distinction.

those who wished to participate in FOAP an important part of confounding bias disappears[51]. Thereby, this second analysis will only consider *treated* individuals with different doses of treatment. Taking into account all these aspects, now we can formally define the unconfoundedness assumption as follows:

Assumption 2b. For each k and $d \in \text{supp}\left(\widetilde{D}_k|\boldsymbol{X}\right)$ and $t \in \{2, \dots, t_T\}$ such that $t \geq g - \delta$,

$$\mathbb{E}[Y(0)_{t-\delta-1}|\boldsymbol{X}] - \mathbb{E}[Y(0)_{t-\delta-1}|\widetilde{D}_k = d, \boldsymbol{X}] = \mathbb{E}[Y(d)_t|\boldsymbol{X}] - \mathbb{E}[Y(d)_t|\widetilde{D}_k = d, \boldsymbol{X}]$$

The plausibility of the unconfoundedness assumptions will be judged with placebo outcome analyses, as in similar studies like Choe et al. (2014). In binary DID strategies, the examination of pretreatment trends through the event-study coefficients of lags may be understood as placebo tests (Eggers et al., 2024). Thereby, we use lagged outcomes in the long past as placebo variables. In continuous DID strategies, the number of coefficients of a regression of the placebo outcome on the treatment variable may be greater than one if we specify a non-linear or non-parametrical functional form. For this reason, we borrow from the literature of matching with continuous treatments the use of the F statistic as a measure of association between continuous treatment variables and a certain placebo variable (Fong et al., 2018). The idea is the same in both cases: if the effect estimate were unconfounded, since we know that a treatment may not affect an outcome that occurred many quarters before, we should not find an association between treatment and the placebo. The contrary would be evidence of an association between the treatment, unobservables and the outcome variable. If such unobservables influenced the probability of treatment and the outcome, we would have confounding.

The second identification assumption is related to overlap or common support. After conditioning on the covariates that make unconfounded the treatment assignment, this assumption asserts that there is enough randomness to find individuals both treated and untreated to estimate counterfactuals. The non-zero probability is a mathematical condition to allow the estimation of conditional

[51] This strategy has also been pursued in other evaluations of adult training programs like Cueto et al. (2010).

probabilities, whereas the non-one probability imposes the existence of chance to find this comparable but not-treated observation. We can attain it through the interpolation or the extrapolation of the data (King & Zeng, 2007). In social sciences, interpolation is usually preferred as shown by the conventional search of the common support. The importance of avoiding extrapolation has usually been based on the high sensitivity of counterfactuals in regions of no overlap (King & Zeng, 2007; Imbens & Xu, 2024). Assumption 3a accommodates this assumption for the time analysis, whilst Assumption 3b do it for the tool analysis.

Assumption 3a. For each $g \in \mathcal{G}$ and $t \in \{2, \dots, t_T\}$, $\Pr[G_g = 1] > 0$ and $\Pr[g, t|\boldsymbol{X}] < 1$ for all $\boldsymbol{X} = \boldsymbol{x}$ in the population of interest.

Assumption 3b. For each $k \in \{1,2,3\}$ and $d \in \text{supp}(\widetilde{D}_k)$, $0 < \Pr\left[\widetilde{D}_k = d \middle| \boldsymbol{X}\right] < 1$ for all $\boldsymbol{X} = \boldsymbol{x}$ in the population of interest.

We will check the plausibility of the overlap condition by examining whether there is common support for all the distribution of the propensity score in each $ATT(g,t)$ (Dehejia & Wahba, 1999). Following Stuart (2010), we will discard those units outside the common support[52]. For the tool-based analysis, as recognized by observational studies with continuous treatments (Huber et al., 2020), completely accomplishing Assumption 3b is a very difficult task. However, we have tried to minimize the consequences of its absence by only considering the subset of individuals that had common support in the binary treatment analysis.

3.2.3. Estimation and inference

For the time analysis, the difference-in-differences strategy is performed in two steps. Firstly, each $ATT(g,t)$ is estimated with the doubly robust estimator of Sant'Anna & Zhao (2020). This estimator models both the treatment process and the outcome process, but only requires the proper specification of at least one of

[52] In this case, the population of interest reflected in the target parameters should be updated with an extra condition: being part of the subset of the population with common support in *all* time horizons. Note that the raw control group, with an assumption on the not-yet-treated individuals, changes at each t. If we only dropped the observations without common support at a certain t, we would generate an unbalanced panel.

these confounding channels. The adjustment on the treatment process is done through inverse probability weighting, estimating the propensity score with a logit model. The outcome process is modelled directly through logistic regression with the same set of covariates introduced in the propensity score equation. Secondly, these group-time effects are aggregated to estimate the target parameters. Inference is performed through the multiplier bootstrap method described in Callaway & Sant'Anna (2021), which has the advantage of producing simultaneous confidence intervals.

For the tool analysis, the use of continuous treatments requires some particularities. To avoid complications related to the staggered implementation in two-way fixed effects models (Callaway, 2023), we only considered those cohorts in which there is not strong evidence of different cohort effects. In practice, this means selecting those cohorts in which the assumption of treatment effect homogeneity seems most plausible. This allows us to pool all the data in only one regression for each e, without the requirement of estimating one equation for each $\{g, t\}$. The first stage is the estimation of the potential outcomes ($ADRF^{(k)}$). Callaway et al. (2024) show the problems of mistakenly assuming a linear effect of the treatment on the outcome. Motivated by their research and by previous empirical results, we make a more flexible specification by using a quadratic functional form. Like the proposal of Callaway et al. (2024), we estimate a model in which the outcome is a first-difference (a trend) formed from the base period. The second stage involves the estimation of the contrasts of interest. We use the {marginaleffects} package to estimate the contrasts and to make inference (Arel-Bundock et al., forthcoming). Standard errors and confidence intervals were estimated analytically through the delta method.

In both cases, as is common in the applied microeconomics literature, we use linear models as "the best linear approximation to a possibly nonlinear conditional expectation function" (Caetano & Callaway, 2023). Wooldridge (2023) has proposed generalized linear models that are robust to the treatment heterogeneity problems of two-way fixed effect specifications. However, its current software implementations have problems to accommodate categorical control variables.

4. Results

This section presents the findings of our study with a focus on the heterogeneity of effects driven by time dimensions and by design features of the tools. Firstly, we leverage the long panel dataset and the staggered implementation of the program to estimate the effects at different horizons and for different cohorts. Secondly, we dig into the design of the program to estimate the effect of each tool.

4.1. Time-based heterogeneity

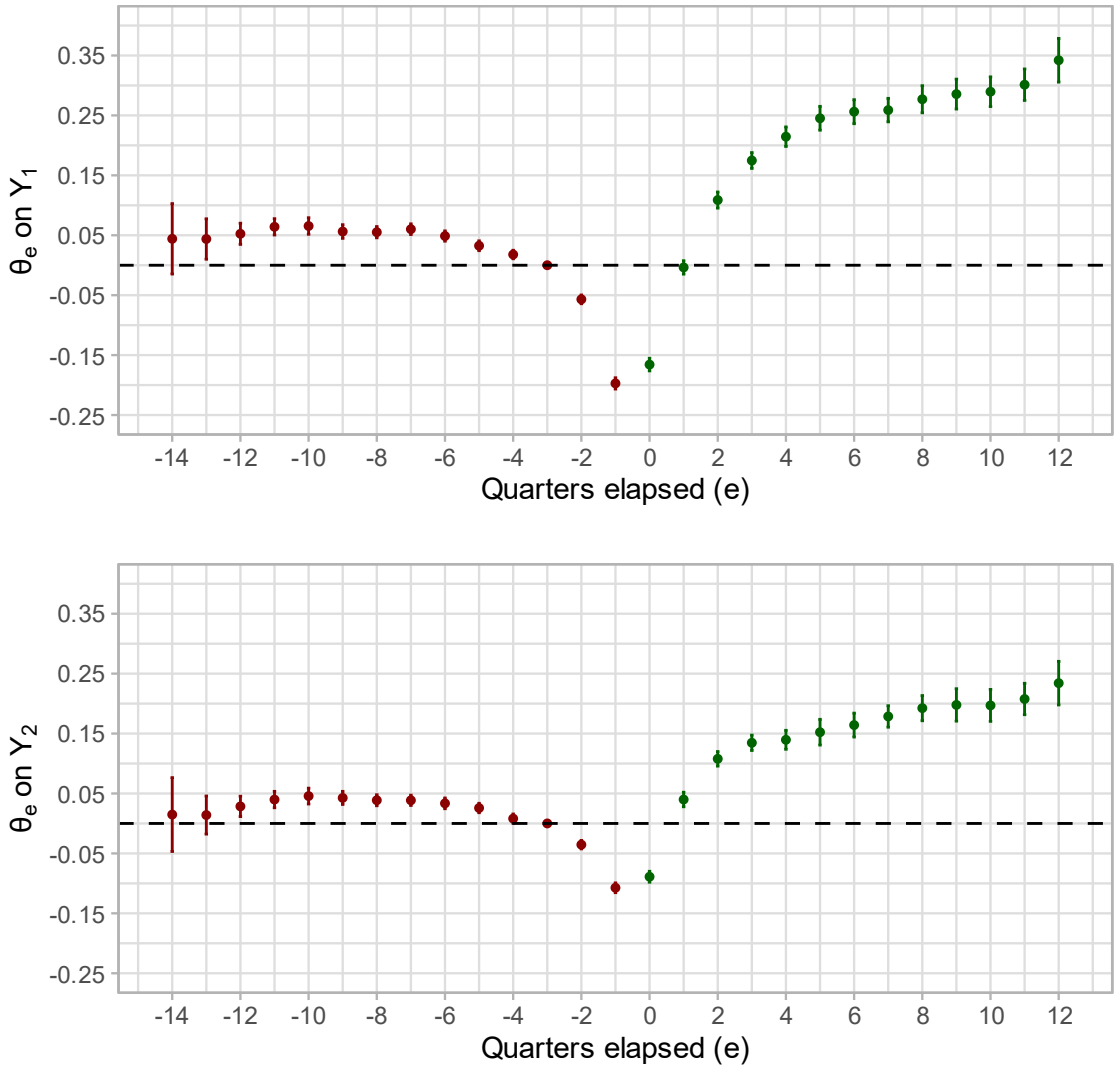
The first part of the results presents the effects in function of the elapsed duration (θ_e). Figure 1 plots the estimates according to the number of quarters since the start of the program (e). Red points denote pre-treatment periods, whereas green ones are post-treatment periods. Concerning post-treatment impacts on the employment probability (Y_1), we detect a lock-in episode of negative and null effects during the first six months. This is a period equivalent to the one reported by Lechner & Wunsch (2009) for trainings in Germany. After this period, the first plot shows that the employment effect is positive and remarkably increasing until quarter five. From this period, the effect is still positive and exceeds the 25 p.p., but its growth slows down. Therefore, we have evidence of dynamic effects by observing that the effect changes in function of the elapsed duration. A similar pattern of decreasing returns was observed in Aragon (Spain) by Clemente et al. (2014), although their estimates converged to approximately 13 p.p. during a recession and to 10 p.p. before it.

These employment gains did not always take the form of open-ended contracts. The second plot of Figure 1 shows the effects on the probability of having an open-ended contract (Y_2). This time the lock-in effect lasts only one quarter, because during months 4 to 6 (quarter 1) we do observe a small but positive effect. From quarter 2, the effect on the probability of open-ended contract keeps positive and increasing. It moves from 10 p.p. to around 20 p.p. in quarter eleven. Therefore, we see how for both outcomes most part of the dynamics of effects occurs during the first year. This finding is again consistent with the stagnation of effects on stable employment detected in Wood et al. (2024), with the estimates of our effect sizes located between the impacts of their classroom-based and their work-based programs.

To complement the visual presentation of findings, Table 1 shows a closer examination of point estimates and standard errors of θ_e for certain time horizons. In the very short run, after six months, we find that the program had a small effect near to zero. It was null for the employment probability and positive for the probability of having an open-ended contract. These results suggest us that during the first half of the year, there might be individuals who suffered a lock-in effect whilst others did find a new job. We will explore later the different duration of the courses as an explanation of this phenomenon. In the short run, after one year, we estimate an effect of 15.2 p.p. on the employment probability. The increase is smaller for the open-ended contract outcome, since we find an impact of 13.5 p.p. for participating in FOAP. The estimate after one year is close to some of the best results for experiments in America (Levy Yeyati et al., 2019).

Looking at the medium and long run, the dynamics of the effects (the second derivative) is less pronounced. This slowdown is clearer in the case of the open-ended contract outcome. Concerning outcome Y_1, after two years (medium run), participating in the program provoked an increase of 26 p.p. in the probability of finding an employment. As we advanced, waiting one additional year rises this effect with only 4 p.p. more. Regarding outcome Y_2, after two years, having participated in FOAP grows the probability of signing an open-ended contract in 18 p.p. One year later, the impact only increases by 3 p.p., but the results are still very noteworthy since most of the new employments take the form of open-ended contracts. All estimates are statistically significant at $\alpha = 0.01$, something that should not be surprising due to the large sample size. Moreover, notice that our estimates of the employment effect at the medium and long run are larger than the 75 % of the sample of non-experimental estimates collected in Card et al. (2018).

Figure 1 – Estimates of θ_e of the participation in FOAP on employment (Y_1) and stable employment (Y_2).



Note: Estimates using a universal base period with anticipation of two periods ($\delta = 2$). It is shown the point estimate and the confidence interval for $\alpha = 0.05$. In all plots, θ_e and $\theta_{g,e}$ must be understood as estimates of the true values.

The isolation of the program effect from other confounders depends on the identification assumptions. In our case, the unconfoundedness assumption is satisfied if bias stability (parallel trends) actually holds. This condition is unverifiable with observational data, but its credibility is usually judged with pre-treatment trends[53]. Red points in Figure 1 report differences in Y_1 and Y_2 between

[53] However, this is neither a sufficient nor a necessary condition for post-treatment parallel trends.

all treatment and control groups in periods before treatment implementation. We took quarter $e = -3$ as the base period, i.e. the moment at which unit fixed effects are estimated. The selection of such period basically considers that $e = -3$ is the period just before the moment at which differences between group change more[54]. To the right of $e = -3$, we confirm the presence of a pattern in adult training programs: the Ashenfelter's dip (Heckman et al., 1999). To be conservative with our estimation, we interpret it as program endogeneity ($Y_{t-1} \to D_t$) and assume that such decrease would be transitory in case we could observe the counterfactual outcome of participants after treatment. To the left of $e = -3$, the plots show the difference between treated and control units in pre-treatment outcomes, which can be understood as placebo outcomes (Eggers et al., 2024). We see that small but statistically significant differences remain between treatment and control groups. Nonetheless, we believe that the CDID assumptions are still plausible due to two reasons. First, adjusting for the confounders mentioned removed a big part of the differences between treated and control units and the remaining differences are small compared with the effect estimates[55]. Second, differences in pre-treatment outcomes are quite constant, which might indicate that the remaining confounders might not have a time-varying effect. If this is the case, we could conduct some analyses assuming that the effect observed before the treatment would replicate after treatment if we could observe the untreated potential outcome.

[54] The selection of the anticipation periods (δ) is explained in more detail at Appendix B.

[55] Appendix D incorporates plots for the raw differences between treatment and control groups.

Table 1 – Estimates of θ_e for conventional horizons on employment (Y_1) and stable employment (Y_2).

	Employment	*Open-ended contract*
$\theta_{e=1}$ (very short term)	-0.004	0.040***
	(0.004)	(0.004)
$\theta_{e=3}$ (short term)	0.175***	0.135***
	(0.005)	(0.004)
$\theta_{e=7}$ (medium term)	0.259***	0.179***
	(0.007)	(0.006)
$\theta_{e=11}$ (long term)	0.301***	0.208***
	(0.006)	(0.009)

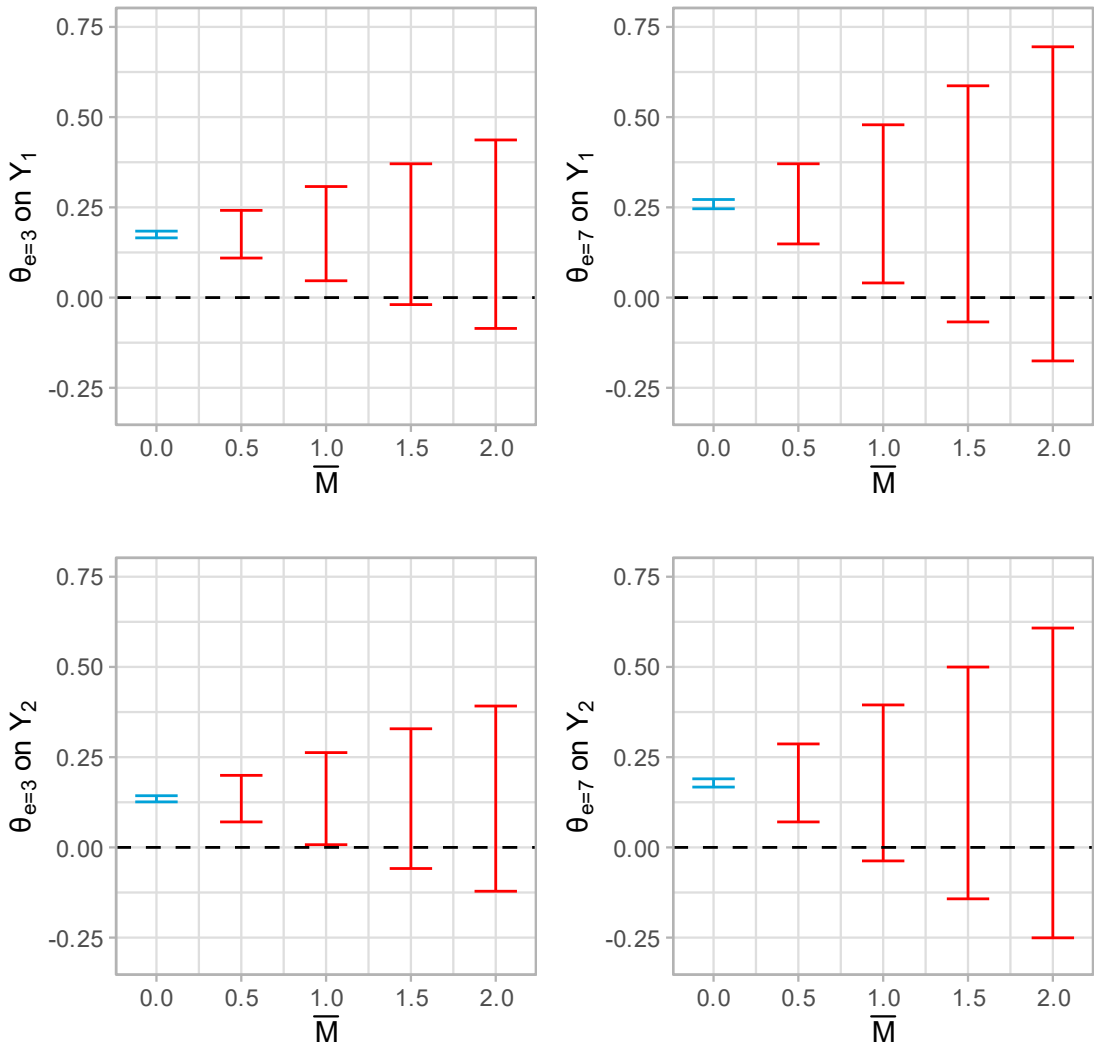
Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. $n_I = 90{,}755$. Each cell shows the point estimate and the standard error in parenthesis. See Appendix C for the rest of estimates.

This is very similar to the proposal of Rambachan & Roth (2023). We leverage our long panel to estimate how deviations from the bias stability assumption would affect the conclusions of our analysis. Since pre-treatment differences do not seem to follow a clear trend, we find more reasonable to experiment with relative magnitudes violations instead of smoothness restrictions[56]. Figure 2 shows how our results might vary if the untreated potential outcome suffered from a confounding bias of size $\overline{M}$-times the greatest difference in pre-treatment outcomes. To simplify the exposition, we only report the analyses for the effects after one ($e = 3$) and two years ($e = 7$). On one side, the effect on the employment probability would still be positive at the short and medium run even if the same (confounding) difference observed in pre-treatment periods was present in the post-treatment period. On the other side, under these same circumstances ($\overline{M} = 1$), the effect on stable employment might lose its positive character. Nevertheless, we should consider that we are talking about confidence intervals and that in this

[56] This last option is very similar to the unit-specific linear trends that some researchers incorporate to address the non-zero pretreatment differences in the outcome (Dustmann et al., 2022). For the same reason, we do not think this option is reasonable in our case.

case the big majority of values whose null hypothesis of zero could be rejected are positive values.

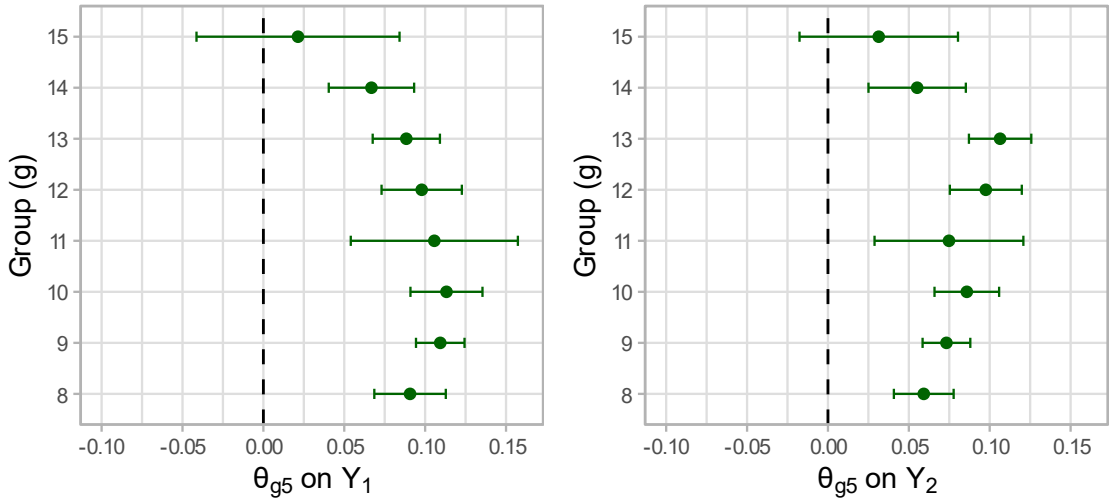
Figure 2 – Rambachan & Roth robust confidence intervals for relative magnitude violations



The second part of the results addresses the cohort effects (θ_{ge}). Figure 3 shows the estimates for each cohort and for each outcome. Regarding the impact on the employment probability (Y_1), the confidence intervals generally overlap and the point estimates are relatively close. This makes us conclude that cohort effects do not seem very important on the effect estimate of the intervention. Concerning the effect on the probability of having an open-ended contract (Y_2), the pattern is similar to the one found for the other outcome with little movements. Removing

the last group due to its large sampling noise, we find that the difference between the biggest and the lowest θ_{g5} is 4.6 p.p. for Y_1 and 5.1 p.p. for Y_2. Although this amount seems small, we must remember that the median effect found by Katz et al. (2022) for sector-focused trainings is 5 p.p. This reminds us that there might be difficulties to estimate these cohort effects related to statistical power.

Figure 3 – Estimates of θ_{g5} of the participation in FOAP on employment (Y_1) and stable employment (Y_2).



Note: Anticipation parameter set to 2. It is shown the point estimate and the 95 % confidence interval. The estimates for cohort 16 were removed due to its small sample size.

If we concentrate on the point estimates, we appreciate a pattern that does not mimic the seasonality effects usually observed in Spanish labour markets (Conde-Ruiz et al., 2019). Consider that $g = 8$ denotes quarter 2016T4 and $g = 15$ denotes quarter 2018T3. Our hypothesis was that those entering the labour market in certain periods might be favoured by the economic situation of this moment. Nonetheless, the terms before the summer that usually concentrate higher levels of new contracts (cohort 9 and 13) do not always show a larger θ_{ge} than the previous quarter. An alternative explanation might be that there are more things changing at each cohort apart from the seasonality surplus. If we estimate the Spearman correlation between θ_{g5} and the adjusted accumulated growth rate, we

get a noteworthy value ($\rho = 0.262$)[57]. Our interpretation is that, although both phenomena seem associated, the effect that might have one on the other does not seem large. We will also explore design explanations in the next section.

Table 2 – Estimates of $\theta_{g,e=5}$ on employment (Y_1) and stable employment (Y_2).

	Employment	*Open-ended contract*
$\theta_{g=8,e=5}$ (2016T4)	0.091***	0.059***
	(0.008)	(0.007)
$\theta_{g=9,e=5}$ (2017T1)	0.109***	0.073***
	(0.006)	(0.005)
$\theta_{g=10,e=5}$ (2017T2)	0.113***	0.086***
	(0.008)	(0.007)
$\theta_{g=11,e=5}$ (2017T3)	0.106***	0.075***
	(0.019)	(0.017)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. $n_I = 90{,}755$. Each cell shows the point estimate and the standard error in parenthesis. The first column reports in parenthesis the specific quarter that denotes each g. See Appendix C for estimates with $g > 11$.

The estimates of these cohort parameters for the maximum number of elapsed periods observed for all cohorts ($e = 5$) are presented in Table 2. The aggregate effects both on employment and on the open-ended contract probability barely change from one cohort to the next. This means that we do not have evidence on remarkable scarring or premium effects attributable to the economic environment of a certain cohort (season). As before, all the coefficients are statistically significant at $\alpha = 0.01$. Therefore, the conclusions we might extract from this analysis are mainly two. On the one hand, if we want to compare two estimates close in time and belonging to the same phase of the business cycle, the part of the effect that might be associated with the cohort is usually small or indistinguishable from sampling error. On the other hand, if the comparison is

[57] We took the calendar and seasonality adjusted quarter GDP growth rate for Spain from INE (2024).

between different stages of the business cycle, the accumulated growth rates differences would be larger, so differences in effect estimates might collect in a remarkable share these cohort effects.

4.2. Tool-based heterogeneity

The second part of the analysis investigates on whether different components of the programs generate different effects. To do that, remind that we have exclusively selected those individuals that participated in FOAP and took advantage of the variation in the dosages of each component. Like in the previous analysis, we start with a visual presentation of the results and then present the quantitative estimates of the contrast parameters.

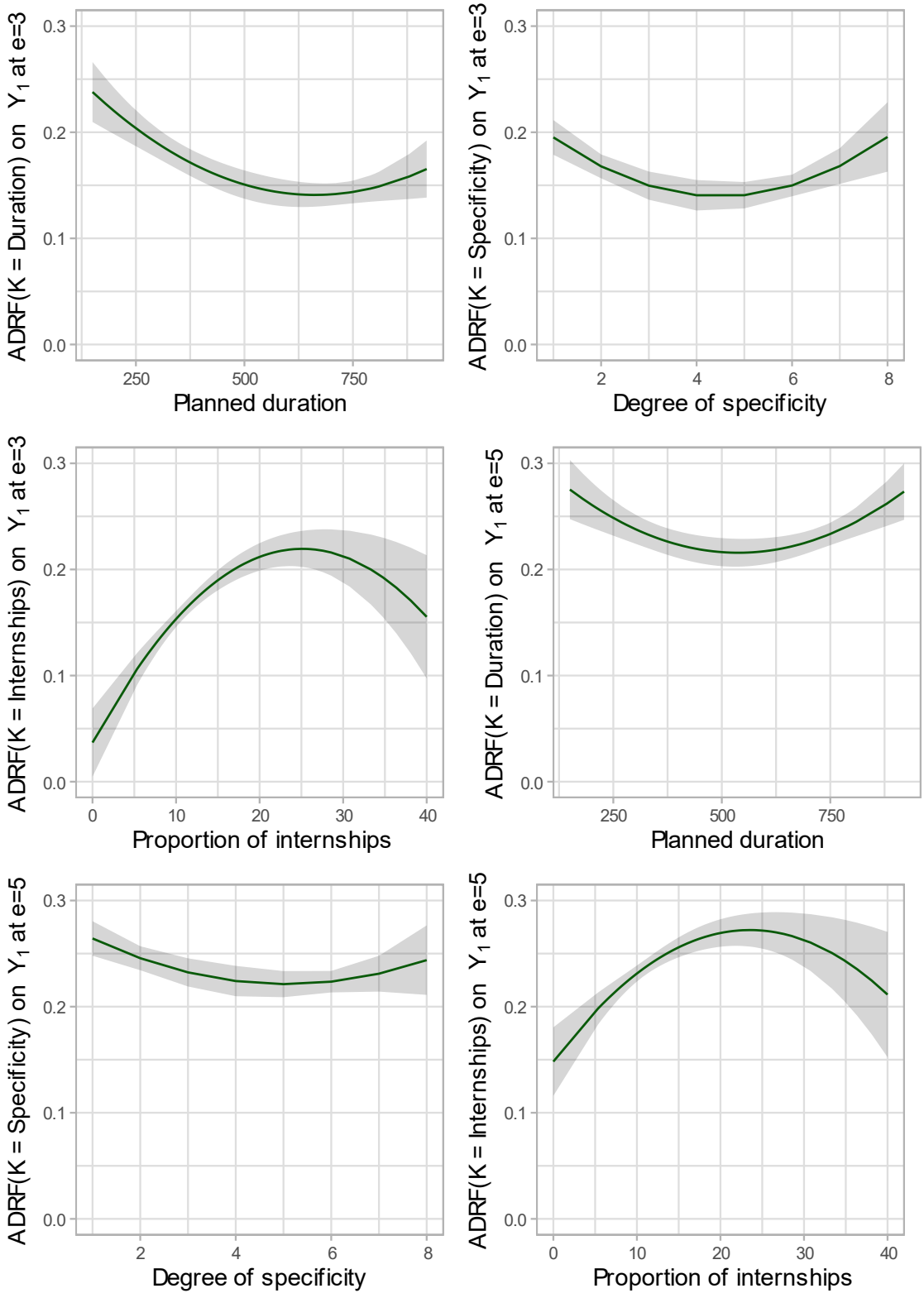
Figure 4 shows the estimated $ADRF_e^k$ about the potential outcome of each component k on the employment probability (Y_1). They can be thought as average predictions of the treated potential outcome for different dosages of each component fixing the value of the rest of design components. We have selected two time horizons observed for all cohorts of interest.

The first and fourth plots regard the *planned duration* of the treatment, measured in hours. The estimates indicate that, controlling for the rest of covariates, the potential outcome is a decreasing function of the planned duration in almost all its domain. It is true that its slopes seem to involve bigger changes one year after treatment start ($e = 3$) than one and a half year ($e = 5$). One year after the beginning of the treatment, our results suggest that the shortest treatments are those that might have the biggest effects on employment. This might be result of the smaller lock-in period of these treatments, since the findings for $e = 5$ qualify the previous statement. After six quarters, the functional form resembles more to a symmetric convex function. The longer treatments gain strength and we see that the treated potential outcome with the highest dosage of duration equals the one obtained after the minimum dosage of duration. This finding is a bit puzzling, suggesting that there might be other design components confounding this result. In any case, our results challenge a big part of the previous literature on the duration of treatments that only focused in one component at each time. This literature usually described this dose-response function as concave (Clemente et al., 2014), although some authors did point up the limited importance of the duration component (Choe et al., 2014).

The second and fifth plots are related to the *occupational specificity* of each treatment, measured through the number of occupations targeted by each training. Our results indicate, for both horizons, that there is a convex relationship between the degree of specificity and the employment probability. We do find that those trainings that are highly specific (targeted at one occupation) present extraordinary levels of employment probability. However, our results also point up that, one year after the beginning of the treatment, the trainings that were lowly specific (targeted at least at eight occupations) also had high levels of employment probability. This situation is softened at $e = 5$. At this time horizon, we estimate a function that gets closer to the direct relationship that is obtained by other researchers of this phenomenon. Still, the convex function for one year after starting the treatment is difficult to understand under the theoretical frameworks discussed in Bol et al. (2019). At least from those trainings aimed at 1 until 6 occupations, we obtain a direct relationship that associates greater specificity with higher employment probability. However, two elements should be taken in consideration when comparing our results with the rest of the literature. Firstly, this literature measures the occupational specificity in a different way, i.e. as realized matches. Secondly, one of the modelling assumptions of DiPrete et al. (2017) and DiPrete & Chaet (2023) is that the dose-response function is a linear function, something that we smooth with a quadratic function.

The third and sixth plots illustrate the *proportion of internships* and its dose-response function related to the employment probability. The key message of these figures is that this is the component with the largest support of the function. In other words, this means that it is the component whose variability is associated with highest changes in levels of employment. This is especially clear if we look at the potential outcome function one year after treatment started. The function is clearly concave, showing the positive effect of internships until the proportion locates around the 25 % of the total duration of the treatment. Still, note that a treatment with the highest possible value in our sample (40 %) has greater employment probability than a treatment which is exclusively classroom-based. These results indicate that the proportion of internships, controlling for the other two components, and in average, seems the design factor that might generate the highest changes in employment. Our findings go in line with Wood et al. (2024) by signalling a remarkable effect of including internships in a program, although they do not look at the impact of different intensities of such component.

Figure 4 – Estimates of $ADRF_e^k$ for employment (Y_1) at one year after treatment started ($e = 3$) and one year and a half ($e = 5$).



Note: Estimates based on the reduced sample formed by those groups without strong evidence of cohort effects, i.e. removing group 15.

Although examining the potential outcomes gives a general view of the dose-response functions, a more quantitative perspective offers more direct results. Table 3 presents point estimates and hypothesis tests for certain contrasts of interest, which are estimates of the target parameter $ACR^{(k)}_{dj,e}$. The specific values selected to form the contrasts are either taken from previous categorizations in the case of the duration (Cueto et al., 2010; Clemente et al., 2014) or from jumps resulted from dividing the dominion of the component in four parts. We are going to focus on the longest horizon to reduce the differences in the lock-in period that involves each dosage of planned duration.

Concerning the planned duration of the treatment, consistent with the previous plots, the shortest and the longest interventions seem more successful than the intermediate ones. In all cases, the differences between the contrasts analysed are statistically significant at the usual levels. If we focus on the occupational specificity of treatments at the longest horizon, we see that reducing the specificity of the treatment in average decreases the employment probability, but only until targeting 4 occupations. From this threshold, we do not find statistically significant differences for the two contrasts presented. Lastly, notice that the column on the proportion of internships offers the largest average causal response. Changing the proportion of internships from 0 to 10 % increases in average the probability of finding an employment in 11.6 p.p. at $e = 3$ and in 8.3 p.p. at $e = 5$. However, notice that this component has not a monotonic effect, since for both horizons going from 30 to 40 % of internships involves a statistically significant reduction in the employment probability.

Table 3 – Estimates of $ACR^{(k)}_{dj,e}$ for employment (Y_1) at one year after treatment started ($e = 3$) and one year and a half ($e = 5$)

	Planned duration	*Specificity*	*Proportion of internships*
$e = 3$			
Δ_1	-0.034***	-0.027***	0.116***
(250-150; 2-1; 10-0)	(0.008)	(0.006)	(0.016)
Δ_2	-0.053***	-0.027***	0.059***
(500-250; 4-2; 20-10)	(0.010)	(0.006)	(0.007)
Δ_3	-0.007	0.009	0.001
(750-500; 6-4; 30-20)	(0.008)	(0.007)	(0.010)
Δ_4	0.022*	0.046***	-0.057***
(920-750; 8-6; 40-30)	(0.012)	(0.016)	(0.020)
$e = 5$			
Δ_1	-0.027***	-0.019***	0.083***
(250-150; 2-1; 10-0)	(0.008)	(0.006)	(0.016)
Δ_2	-0.032***	-0.022***	0.038***
(500-250; 4-2; 20-10)	(0.010)	(0.006)	(0.007)
Δ_3	0.017**	-0.001	-0.007
(750-500; 6-4; 30-20)	(0.008)	(0.007)	(0.011)
Δ_4	0.040***	0.020	-0.051**
(920-750; 8-6; 40-30)	(0.012)	(0.016)	(0.020)

Note: The value k of the target parameter is shown in the columns, value e names each panel and the low (d) and high (j) doses of each comparison are shown in parenthesis at the first column sequentially for each k. Each cell presents the point estimate and the standard error in parenthesis, which are estimated analytically with the HC3 correction. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

To assess the credibility of our causal models, we have performed some placebo tests for the three groups of coefficients of interest. Table 4 includes one row for each placebo outcome. Remind that, if our identifying restrictions hold true, the placebo outcome should not be associated with each group of coefficients because it took place before such treatment was implemented. What we appreciate in all cells is that the p-values of all hypothesis tests are over the usual significance level of 10 %. This is good news, since it means that we cannot reject that the true

relationship between each placebo outcome and the set of coefficients of each design component was zero. This is the same type of analysis that has been presented in the previous section by analysing the pre-treatment coefficients at the event-study plot.

Table 4 – Placebo analyses using lagged outcomes of employment (Y_1)

	Planned duration	*Specificity*	*Proportion of internships*
$Y_{1e=-4}$	2.002	0.553	1.054
	(0.135)	(0.575)	(0.349)
$Y_{1e=-5}$	1.959	1.173	0.011
	(0.141)	(0.31)	(0.989)
$Y_{1e=-6}$	0.786	1.507	0.233
	(0.456)	(0.222)	(0.792)
$Y_{1e=-7}$	0.582	0.877	0.174
	(0.559)	(0.416)	(0.84)

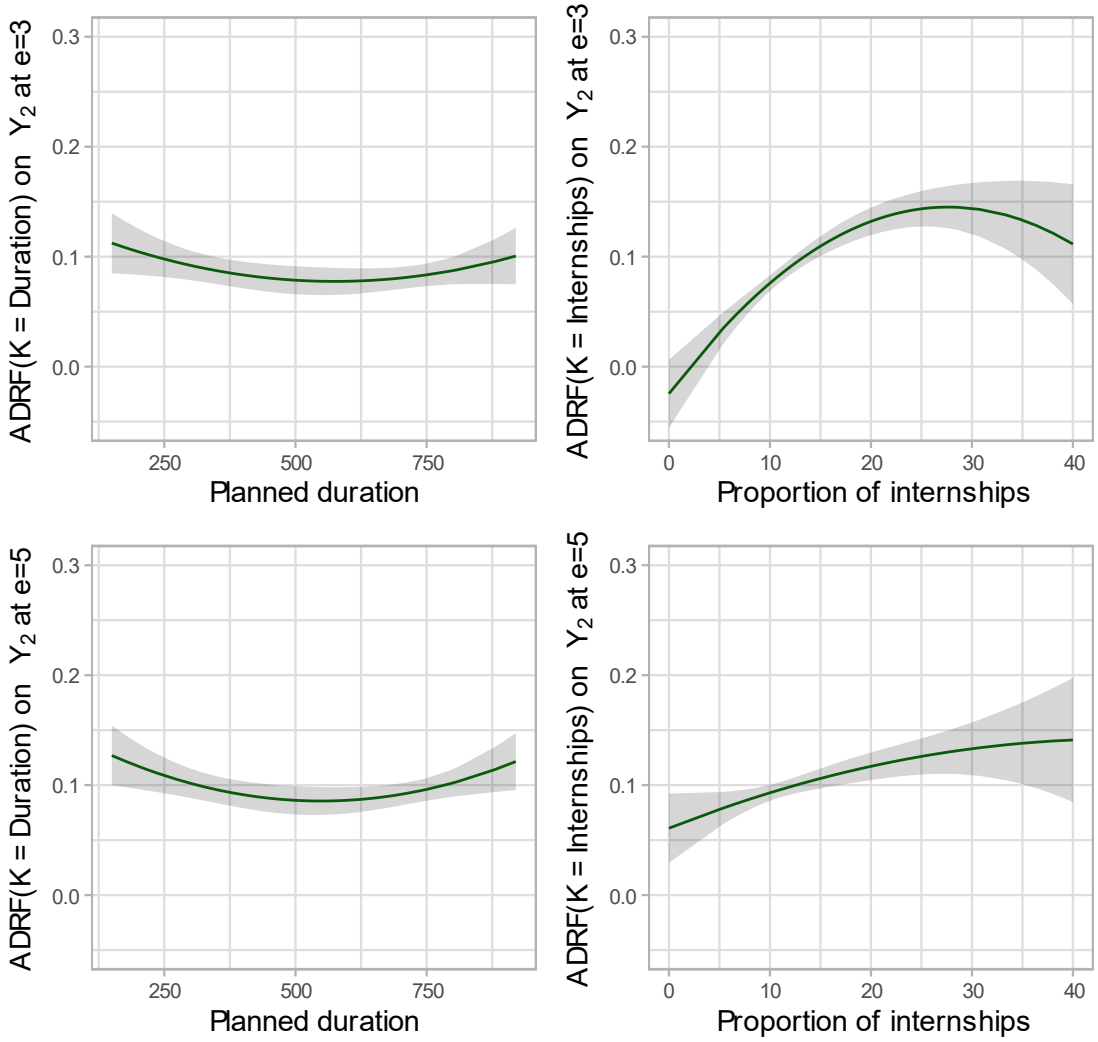
Note: Each cell reports the heteroskedasticity-robust F statistic and the p-value for the null hypothesis of coefficients of the focal component equal to zero in parenthesis.

We can repeat the previous analysis but focusing us on the other outcome, the probability of having an open-ended contract (Y_2). Notice that Figure 5 has not plots on the estimates of $ADRF_e^k$ for the specificity component. The reason is that, as we will see later, we did not pass the placebo tests for this design component. This subtracts credibility to the causal model of the effect $\widetilde{D}_{k=2} \rightarrow Y_2$ assumed in the main analysis, so we have opted for removing the results of their estimates. In any case, we do have credible estimates for the effect of the duration and the proportion of internships on the probability of stable employment.

In Figure 5, the two plots on the average dose-response function of the duration component on the probability of stable employment (Y_2) are not very promising. The function is almost constant around the 10 p.p. of stable employment probability, without remarkable changes after varying the planned duration of the treatment. One year and a half since the beginning of the treatment ($e = 5$), the

function slightly gets greater upward curvature, which would indicate better results when this component has extreme values.

Figure 5 – Estimates of $ADRF_e^k$ for stable employment (Y_2) at one year after treatment started ($e = 3$) and one year and a half ($e = 5$).



Note: Estimates based on the reduced sample formed by those groups without strong evidence of cohort effects, i.e. removing group 15.

The other two plots of Figure 5 deal with the average potential outcome of stable employment in function of the proportion of internships. The $ADRF_e^k$ for $e = 3$ has a similar functional form to that found for Y_1: the more proportion of internships, the greater is the potential outcome until a point around the 25 %. However, the relationship turns almost linear when we observe the outcome two

quarters later. Although there is more sampling noise for the values obtained with extreme values of internships, it seems that at $e = 5$ there is an increasing average dose-response function. Wood et al. (2024) did not study different effects for different intensities, but they also found more probability of stable employment for programs with more than 0 % of internships.

Table 5 – Estimates of $ACR^{(k)}_{dj,e}$ for stable employment (Y_2) at one year after treatment started ($e = 3$) and one year and a half ($e = 5$).

	Planned duration	*Proportion of internships*
$e = 3$		
Δ_1	-0.014**	0.100***
(250-150; 10-0)	(0.007)	(0.015)
Δ_2	-0.019*	0.056***
(500-250; 20-10)	(0.010)	(0.007)
Δ_3	0.005	0.012
(750-500; 30-20)	(0.007)	(0.010)
Δ_4	0.017	-0.032*
(920-750; 40-30)	(0.011)	(0.019)
$e = 5$		
Δ_1	-0.018**	0.032**
(250-150; 10-0)	(0.007)	(0.015)
Δ_2	-0.023**	0.024***
(500-250; 20-10)	(0.010)	(0.007)
Δ_3	0.010	0.016
(750-500; 30-20)	(0.008)	(0.010)
Δ_4	0.025**	0.008
(920-750; 40-30)	(0.011)	(0.019)

Note: The value k of the target parameter is shown in the columns, value e names each panel and the low (d) and high (j) doses of each comparison are shown in parenthesis at the first column sequentially for each k. Each cell presents the point estimate and the standard error in parenthesis, which are estimated analytically with the HC3 correction. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Since researchers and practitioners are usually interested in contrasts of potential outcomes, Table 5 shows the point estimates of some differences and their statistical significance. The specific contrasts take the same values of the components that were used for the analysis of the employment probability. The main messages are similar to those obtained from that analysis. For the planned duration, we see that the average causal response of going from 150 to 250 hours or from 250 to 500 hours involves a decrease in the probability of having an open-ended contract. This means that shorter programs are generally preferable. For the proportion of internships, we appreciate a remarkable benefit of moving from a 0 to a 10 % internships program especially at the short run, i.e. at $e = 3$. Unlike the employment outcome, if we concentrate on stable employment, there are null effects of increasing the proportion of internships from the 20 % at the usual significance level of 5 %.

Lastly, Table 6 reports the placebo analyses with pretreatment outcomes to add or subtract credibility to an effect estimation. The two components analysed, duration and internships, present hypothesis tests that do not allow to reject the null hypothesis of zero association. This would add credibility to the causal interpretation of our statistical models. As we anticipated before, three out of the four placebo outcomes for the occupational specificity component are statistically significant at usual levels. The conventional interpretation of this phenomenon is that there might be a time-invariant unobserved variable influencing both treatment dosage and the outcome of interest. Since this would involve confounding, we have discarded that analysis.

Table 6 – Placebo analyses using lagged outcomes of stable employment (Y_2)

	Planned duration	*Specificity*	*Proportion of internships*
$Y_{2e=-4}$	1.422	2.186	1.322
	(0.241)	(0.112)	(0.267)
$Y_{2e=-5}$	1.702	3.002	0.415
	(0.182)	(0.05)	(0.661)
$Y_{2e=-6}$	1.181	3.573	0.638
	(0.307)	(0.028)	(0.529)
$Y_{2e=-7}$	0.67	3.7	0.456
	(0.512)	(0.025)	(0.634)

Note: Each cell reports the heteroskedasticity-robust F statistic and the p-value for the null hypothesis of coefficients of the focal component equal to zero in parenthesis.

4.3. Robustness analyses

The credibility of the results presented in the previous sections rests on the plausibility of the identifying restrictions. However, the fundamental problem of causal inference involves identification uncertainty even in DID analyses (Rambachan & Roth, 2023): having evidence of bias stability before treatment does not imply that bias will stay stable after treatment, so we must rest on assumptions. In this section, we estimate how robust are the results by varying two aspects of the main analysis: the assumption on the Ashenfelter's dip and the identification strategy.

On the Ashenfelter's dip, one assumption considers that the dip is transitory, which means that the way to proxy the stable bias affecting $Y\ 0\ _t$ is to account for stable differences before the dip (in our case, $\delta = 2$). The other assumption establishes that such dip is reflecting a scar that would be persistent, at least for the t's of interest. Thereby, we should control for the value of such dip by proxying the stable bias using as base period the difference in the last pretreatment outcome (then $\delta = 0$). On the identification, we have experimented with two non-nested causal models on the untreated potential outcome: the conditional difference-in-differences model (CDID) and the conditional lagged dependent variable model (CLDV). The DID model assumes that confounding bias occurs

due to unit *time-invariant* differences supplemented by a temporal trend (stable bias). Its conditional version recognizes the necessity of accounting for time-varying effects of observed variables to completely remove confounding (Zeldow & Hatfield, 2021). The LDV model asserts that confounding bias happens due to unit *time-variant* differences (unstable bias). Its conditional version also acknowledges the necessity of accounting for the effects of observed variables to completely remove confounding. This sensitivity approach has been suggested in methodological works like Angrist & Pischke (2009), O'Neill et al. (2016), or Ding & Li (2019).

We are going to start by evaluating how sensitive are the results that analysed the effects in function of the period of observation (dynamics). Table 7 shows the results after redoing the analyses related to θ_e. Fixing the identification strategy to the one used in the main analysis, the results shown for the CDID with $\delta = 0$ are too far from the 5 p.p., which is the median impact on employment found at $e = 7$ in the experiments analysed by Katz et al. (2022). This reinforces our idea of fixing δ at 2 instead of the standard value of 0. Importantly, the decision on δ must also be made with the other identification strategy. We should think which assumption on this parameter makes more sense according to a certain model of the untreated potential outcome. If we believe that confounding comes exclusively from time-variant differences between units, we argue that it makes more sense to include all the available pre-treatment lags. The CDID can only consider *one* pretreatment outcome to estimate the unit fixed effect in a first-difference model, but the CLDV may include the terms of the dip even if they actually do not have an effect on the counterfactual untreated potential outcomes. Moreover, O'Neill et al. (2016) show that the problems of serial correlation decrease when the number of lags included increases. Therefore, we focus for CLDV on the results with $\delta = 0$.

Looking at the results from the CLDV identification strategy, Table 7 shows that participating in FOAP in average rises in 7.8 p.p. the employment probability after one year ($e = 3$). It is interesting that, after two years, $\theta_{e=7}$ decreases to 4.5 p.p. when estimated using CLDV, whereas the CDID suggests that the effect goes up. In any case, we think that the bounds established for both approaches are quite informative for the first year of implementation. If the true effect was positive, following Ding & Li (2019), we can assert that the true quantity would be between 7.8 p.p. (the lower bound) and 17.5 p.p. (the upper bound). If a person

was treated, the CLDV model estimates that in average her/his probability of having an open-ended contract increased by 5.8 p.p. at the end of the first year. The effect on this outcome also declines one year later under the CLDV identification. In fact, in the Appendix C we include the complete list of estimates for each e and show that the presence of dynamics in the effect estimates is quite moderate. Considering the international literature, the effect sizes obtained with the CLDV model are more in line with the experimental and the non-experimental results (Levy Yeyati et al., 2019; Card et al., 2018).

Table 7 – Sensitivity analyses of θ_e by outcome, identification strategy and anticipation parameter.

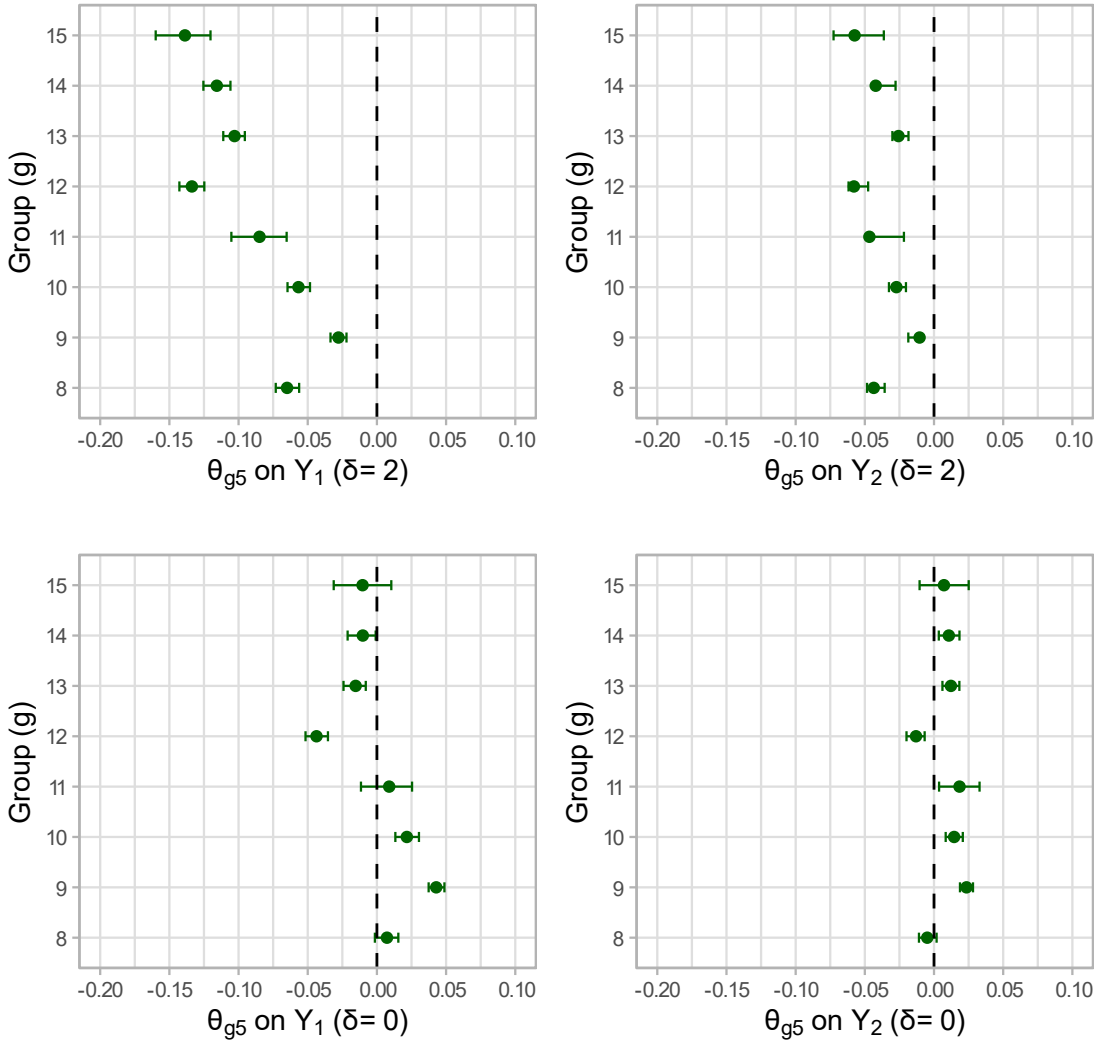
	$\delta = 2$		$\delta = 0$	
	CDID	CLDV	CDID	CLDV
Y_1 *(employment)*				
$\theta_{e=3}$	0.175***	-0.006**	0.382***	0.078***
	(0.005)	(0.003)	(0.004)	(0.003)
$\theta_{e=7}$	0.259***	-0.009**	0.445***	0.045***
	(0.006)	(0.004)	(0.006)	(0.004)
Y_2 *(stable employment)*				
$\theta_{e=3}$	0.135***	0.018***	0.244***	0.057***
	(0.004)	(0.003)	(0.004)	(0.003)
$\theta_{e=7}$	0.179***	0.009***	0.276***	0.032***
	(0.006)	(0.004)	(0.005)	(0.004)

Note: Each cell shows the point estimate and the standard error in parentheses. Inference statistics for CLDV were estimated through nonparametric bootstrap with 500 replications. θ_e's with CLDV have been calculated through the aggregation of their building-block parameters $ATT(g,t)$ following the philosophy proposed in Callaway & Sant'Anna (2021). See Appendix C for the rest of estimates. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

The next sensitivity analyses deal with the estimates of the parameter that captures the cohort effects, i.e. θ_{g5}. To save space we only include the results for the CLDV models. In this case, we do observe the seasonal effects that we missed

in the CDID analysis: those groups that started the treatment before the summer ($g = 9$, $g = 13$) experimented a higher θ_{g5} than the cohort just before. This might suggest not to select the same anticipation parameter for all $ATT(g,t)$'s, since with a universal δ some base periods might also collect differences attributable to seasonality and not only unit time-invariant differences. In any case, if we concentrate us on the findings with $\delta = 0$, we appreciate that point estimates are closer. This is especially clear for the stable employment outcome, in which we only detect a small difference for those periods before those affected by seasonality. This closeness reinforces the idea on the low importance of cohorts.

Figure 6 – Estimates of θ_{g5} using a CLDV identification by outcome and δ.



Note: Confidence intervals were estimated through nonparametric bootstrap with 500 replications.

To finish, we have also estimated how sensitive are the results of the design components analysis when changing the anticipation parameter or the identification strategy. Tables 8 and 9 illustrate that, in this case, the results are practically identical even if we vary the value of δ. These estimates are also almost equal to the ones obtained with CDID, which is good news. The less sensitivity of this analysis to the identification or the anticipation parameter might come from the exclusive use of treated individuals. This might suggest us that the differences between individuals assigned to each dosage might be modest, since the estimation of causal estimands does not seem to change in function of certain statistical adjustments. Again, the findings with CLDV identification in Tables 8 and 9 highlight the positive employment effects of treatments with short duration, very specific, and with intermediate proportion of internships.

Table 8 – Estimates of $ACR^{(k)}_{d,j,e}$ using a CLDV identification for employment (Y_1) at one year after treatment started ($e = 3$) for $\delta = 2$.

	Planned duration	*Specificity*	*Proportion of internships*
CLDV and $\delta = 2$			
Δ_1	-0.033***	-0.023***	0.097***
(250-150; 2-1; 10-0)	(0.006)	(0.005)	(0.013)
Δ_2	-0.052***	-0.024***	0.042***
(500-250; 4-2; 20-10)	(0.008)	(0.005)	(0.006)
Δ_3	-0.009	0.005	-0.013
(750-500; 6-4; 30-20)	(0.006)	(0.006)	(0.009)
Δ_4	0.019**	0.034***	-0.067***
(920-750; 8-6; 40-30)	(0.009)	(0.013)	(0.017)
Passed the placebo test	Yes	Yes	Yes

Note: Since the CLDV strategy requires the inclusion of lagged outcomes in the specification of the model, we only reserved $Y_{ie=7}$ as placebo outcome. For the sake of brevity, we only report the sensitivity analyses for the effects on $Y_{ie=3}$.

Table 9 – Estimates of $ACR^{(k)}_{dj,e}$ using a CLDV identification for employment (Y_1) at one year after treatment started ($e = 3$) for $\delta = 0$.

	Planned duration	*Specificity*	*Proportion of internships*
CLDV and $\delta = 0$			
Δ_1	-0.033***	-0.022***	0.097***
(250-150; 2-1; 10-0)	(0.006)	(0.005)	(0.013)
Δ_2	-0.052***	-0.023***	0.042***
(500-250; 4-2; 20-10)	(0.008)	(0.005)	(0.006)
Δ_3	-0.008	0.006	-0.013
(750-500; 6-4; 30-20)	(0.006)	(0.005)	(0.009)
Δ_4	0.020**	0.035***	-0.067***
(920-750; 8-6; 40-30)	(0.009)	(0.013)	(0.017)
Passed the placebo test	Yes	Yes	Yes

Note: Since the CLDV strategy requires the inclusion of lagged outcomes in the specification of the model, we only reserved $Y_{ie=7}$ as placebo outcome. For the sake of brevity, we only report the sensitivity analyses for the effects on $Y_{ie=3}$.

Regarding employment quality, Table 10 presents the estimates of the effects of the three design components on the probability of open-ended contract but using a CLDV strategy. As happened with the employment quantity outcome, the results do not substantially change neither when altering the anticipation parameter nor when using a different identification strategy. Like with the employment outcome, our models for employment quality suggest focusing on trainings that are short, very specific, and have intermediate proportions of internships. These conclusions are shared both from a CDID and from a CLDV strategy, which reinforces the internal validity of the results.

Table 10 – Estimates of $ACR^{(k)}_{dj,e}$ using a CLDV identification for stable employment (Y_2) at one year after treatment started ($e = 3$) in function of δ.

	Planned duration	*Specificity*	*Proportion of internships*
CLDV and $\delta = 2$			
Δ_1	-0.015***	-0.020***	0.075***
(250-150; 2-1; 10-0)	(0.006)	(0.005)	(0.012)
Δ_2	-0.021***	-0.019***	0.035***
(500-250; 4-2; 20-10)	(0.008)	(0.005)	(0.006)
Δ_3	0.004	0.008	-0.006
(750-500; 6-4; 30-20)	(0.006)	(0.005)	(0.008)
Δ_4	0.016*	0.035***	-0.046***
(920-750; 8-6; 40-30)	(0.009)	(0.012)	(0.016)
Passed the placebo test	Yes	Yes	Yes
CLDV and $\delta = 0$			
Δ_1	-0.015***	-0.019***	0.076***
(250-150; 2-1; 10-0)	(0.006)	(0.005)	(0.012)
Δ_2	-0.021***	-0.018***	0.035***
(500-250; 4-2; 20-10)	(0.008)	(0.005)	(0.006)
Δ_3	0.004	0.009*	-0.006
(750-500; 6-4; 30-20)	(0.006)	(0.005)	(0.008)
Δ_4	0.017*	0.036***	-0.047***
(920-750; 8-6; 40-30)	(0.009)	(0.012)	(0.015)
Passed the placebo test	Yes	Yes	Yes

Note: Since the CLDV strategy requires the inclusion of lagged outcomes in the specification of the model, we only reserved $Y_{ie=7}$ as placebo outcome. For the sake of brevity, we only report the sensitivity analyses for the effects on $Y_{ie=3}$.

5. Conclusions

In this chapter, we have evaluated an adult training program that allows us to estimate the effects of its interventions in function of time dimensions (the period of observation or the cohort) and program design dimensions (the duration,

occupational specificity, or internships proportion). We have focused on two outcomes: the probability of employment and the probability of having an open-ended contract. On the time dimensions, we have detected that the effects on both outcomes are positive and increase over time, although with diminishing returns. After one year, we estimate that the impact is of 17.5 – 7.8 p.p. on employment and of 13.5 – 4.5 p.p. on stable employment if we consider both identification strategies. There is less evidence on a remarkable cohort effect, especially if we compare cohorts whose starting quarters are close. On the design dimensions, our results indicate noteworthy nonlinear effects of the proportion of internships on the employment probability. After six quarters, rising the value of this variable may substantially increase almost linearly the stable employment probability. The impact of the other two components, duration and specificity, are remarkably lower.

We have supported and qualified previous research with our work. The literature had also detected dynamic effects in continental Europe or with an alternative identification strategy. We have confirmed that this pattern is also present in Spain with diminishing returns after six quarters in both outcomes. Still, the lower bound of the effect estimates is higher than most estimates found in the literature (Vooren et al., 2019). The low importance of cohort effects that was estimated in Germany by Lechner & Wunsch (2009) was also found in our case for our program. However, the most innovative part of our research is the one that offers the most challenging results. Unlike previous research who analysed them separately, we tried to control for the specific effects of treatment duration, internships and occupational specificity. We found that internships are the design component that matters most, especially when jumping from zero to low proportions of internships. When controlling for this component, the duration has a convex dose-response function on employment with moderate effect sizes. However, in line with previous results, we also detected the positive effects of very occupational specific treatments. Notice that these design results are very robust to using a CLDV instead of a CDID identification strategy.

Our study also suffers from some limitations. The first caveat is that all effect estimates rest on the validity of the causal models. Although their assumptions are untestable, we have enhanced the specification of causal models with placebo tests and then restricting the analysis population to those treated in any version of FOAP. However, we must recognize that the identification assumptions in

observational studies, and especially if there are continuous treatments, are strong. Continuous treatment designs involve the generation of an infinite number of treatment groups, so the exchangeability and positivity assumptions turn more complicated. Due to these difficulties and the particularities of a compound treatment, we recommend policymakers to consider a factorial experimental design with a small subset of participants (Wu & Hamada, 2021; Rubin, 2008). The second warning is related to our measurements of employment. The available datasets only allow to observe if the individual was working at the last day of the quarter and new labour contracts signed at any date. We do not know the actual duration of each labour relationship, so this must be considered when interpreting our results thinking in certain precarious jobs. The third caveat regards the transportability of our findings to other less developed regions of Spain. We have concentrated us on a regional economy with lower natural rates of unemployment than other regions like Andalusia or Extremadura (Cuéllar-Martín et al., 2019). Directly generalizing our findings to these contexts means neglecting the effect modifier abilities of supply-side factors like productive structure, which is an issue that deserves a closer attention (Pearl & Bareinboim, 2014).

Future research might also explore alternative paths even if we continue with a non-randomized treatment assignment. Maintaining the same research questions, on one side, researchers might use a third causal model of untreated potential outcomes: the interactive fixed effect model (Callaway & Karami, 2023; Xu, 2017). This structural model allows the effect of the confounders to vary over time, but comes at the cost of additional requirements, so analysts might evaluate its plausibility in further works. On the other side, another line of research could focus not on data analysis but on data collection. There is space for measuring the actual occupational specificity of treatments à la DiPrete et al. (2017) instead of the perceived specificity that we have exploited. More generally, we think that it would be valuable to develop coding instructions to transform the text of design documents into a more detailed and standardized classification of training programs. This would involve more design components that the ones we have considered here, as it has been proposed in Junquera (2024b) for employment subsidy programs. Slightly varying the research questions, future evaluations might explore different outcomes or different effect modifiers. Models might incorporate theoretical mediators of employment like job search self-efficacy as dependent variables (Liu et al., 2014), but also concentrate on features of the

population of participants as effect modifiers. Researchers might employ random forests to understand deviant groups of participants with high rates of post-treatment employment and then interview them to search new covariates that might explain such exceptional cases (Harding & Seefeldt, 2013). A completely different research question would be a cost-benefit analysis, which would also be quite interesting due to their low use in the ALMP literature of Southern Europe.

All these warnings and research gaps help us to understand that the impact evaluation of adult training is an ongoing and incomplete task. Nonetheless, our results indicate that the employment effects of these programs in Spain seem positive and illuminate promising design components like internships and occupational specificity.

Appendix

Appendix A. On the variables

In this section, we define how we measured each variable and indicate its level of measurement in parentheses.

A.1. Outcome variables

Employment at quarter *t* (qualitative). We are measuring if the individual worked at least one day during such quarter. An individual is coded as employed (employee or self-employed) at quarter t if (a) he was affiliated as employee or self-employed to the Social Security during the last day of such quarter or (b) a labour contract was registered at any day of such quarter.

Open-ended contract at quarter *t* (qualitative). This variable collects if the individual had an open-ended contract at the last day of the quarter (measured through the Social Security dataset) or if he/she signed an open-ended contract at any day of the quarter (measured through the Contracts dataset).

A.2. Treatment variables

Duration of the treatment (quantitative). Taking the planned duration avoids the self-selection problems related to considering actual duration of the treatment. While the planned duration is not decided by the participant, the actual duration it is. The literature has categorized this variable in different ways. Table A.2.1. shows the different categorizations of Cueto et al. (2010) and Clemente et al. (2014), two of the available evaluations of adult training in Spain. The program design of FOAP establishes the number of hours that must last a certain treatment, but not the specific calendarization that must follow. In other words, it does not determine the number of hours per week or per day, only a maximum (40 hours per week). Table A.2.2. approximates the number of weeks and quarters that would entail a full-time treatment (40 hours per week) for an operative categorization of hours.

Table A.2.1. – Categorizations of the duration variable in other studies

Study	*Category*	*Interval*
Cueto et al. (2010)	Short duration	hours ≤ 250
	Intermediate duration	$251 \leq$ hours ≤ 500
	Long duration	hours > 500
Clemente et al. (2014)	Short duration	hours < 150
	Intermediate duration	$150 \leq$ hours ≤ 300
	Long duration	hours > 300

Table A.2.2. – Frequency of treatments by planned duration in hours for FOAP

Category	*Hours*	*Weeks for a FTT*	*Quarters for a FTT*
Short	$h \leq 250$	$3.75 \leq w \leq 6.25$	$0.28 \leq q \leq 0.48$
Intermediate	$250 < h \leq 500$	$6.25 < w \leq 12.5$	$0.48 < q \leq 0.96$
Long	$500 < h \leq 750$	$12.5 < w \leq 18.75$	$0.96 < q \leq 1.44$
Very long	$h > 750$	$18.75 < w \leq 23$	$1.44 < q \leq 1.77$

Note: FTT: Full-Time Treatment (40 hours per week). min $h = 150$, max $h = 920$.

Occupational specificity (quantitative). Following the review by Blommaert et al. (2020), we see that the literature on training has used a dichotomic vocational-general variable to characterize educational systems. We think this phenomenon as *one* variable (dimension) measured through two polar (extreme) values: vocational (no general) versus general (no vocational). Digging into this approach, we follow two changes to such operationalization inspired by DiPrete et al. (2017). Firstly, to measure it continuously. Secondly, to apply it to credentials and not to educational systems (sets of credentials). We make a distinction between actual specificity and perceived specificity. While actual specificity operationalizations focus on realized employment matches (Bol et al. (2019)), perceived specificity is a property of the program design as long as it is defined through the number of occupations targeted by the designers. Based on a sociological tradition interested on the unintended consequences of actions (Linares Martínez, 2018), we think that it is more useful to concentrate on the *perceived* occupational specificity of a credential. To operationalize this variable in our study, we count the number of CNO-11 occupations (at 4 digits) related to each credential. In our dataset, there are 154 different credentials (in Spanish, *especialidades formativas*). With the

automatic scraping, we got information on the linked occupations for 133 credentials. We could also collect information manually on 9 credentials by exploring their regulation. We assign them the CNO occupation closes to the literals found thanks to the assistance of AYUDACOD, a search tool developed by INE. This makes a total of 142 credentials with information. The remaining 12 credentials with unknown match to occupations were all from the family of languages. In terms of specificity, these credentials are highly general (lowly specific). Thereby, we impute them the lowest value for this variable.

Internships (quantitative). According to their degree of regulation, there are two types of treatments: those that grant a professional certificate (CP) and those that do not (non-CP). For CP treatments, we took the planned duration of the internships from Dades Obertes Catalunya (2023). In this case, considering the planned and not the actual duration has also the benefit of avoiding confounding bias (self-selection). Notice that the individual may extend the actual length of the internships through the so-called *prácticas de perfeccionamiento* (improvement internships). However, we have only considered the compulsory duration of the internships. For non-CP treatments, the planned duration of the internships may vary from provider to provider[58], and it is not available in our datasets. Therefore, we considered the actual duration of internships. Unlike CP treatments, improvement internships in non-CP treatments are not a decision made exclusively by the participant. To carry out these internships, the provider must ensure that they will be pursued by at least 60 % of the treatment participants. This ameliorates the suspicions on confounding due to unobserved individual covariates when internships are registered. Notice that to make this variable unsensitive to the planned duration of a treatment, we focus on the *proportion* of hours of internships.

[58] Unless they were mandatory by regulation.

A.3. Confounding variables

Table A.3.1 – List of confounders or proxies of confounders.

Vector	*Variable*	*N. of categories*	*Adjusted in time analysis*	*Adjusted in tool analysis*
Socio-demographic ($\boldsymbol{S}_i$)	Sex	2	X	X
	Age	∞	X	X
	National group	5	X	X
	Level of education	6	X	X
	Disability	2	X	X
Labour market ($\boldsymbol{L}_i$)	Local labour market	14	X	X*
	Job search area	7	X	X
	Days with benefit in the year before treatment	366		X
Proxies of latent confounders ($\boldsymbol{U}_{it}$)	Unit component	n	X	
	Time component	t	X	X
	Lagged dependent variables	2		X

Note: (*) Due to the high number of treatment groups in the design analysis, we had to collapse some levels of the local labour market variable to avoid empty or tiny sample sizes. The resulting variable has 14 levels.

Sex (qualitative). We considered two categories: woman and man.

Age (quantitative). This variable measures the age in years as of 1 January, 2017. We introduce it with a quadratic functional form in all specifications to have more flexible potential outcome models.

National group (qualitative). We substantially followed the experimental results of Polavieja et al. (2023) to group individuals into national groups. Thereby, their results indicate modest but statistically significant discrimination towards two regions of ancestry in Spain: (1) Middle East and North Africa, and (2) Latin

America and the Caribbean. They did not find a statistically difference for those with ancestries from Asia versus Europe-USA. Although the study did not test for differences between citizens and not citizens of the European Union, labour regulation is remarkably stricter for the second group. Although possibly they are not discriminated in the labour market, it seems plausible to assume that non-Spaniards EU citizens had a lower probability of FOAP participation due to language issues. Therefore, we finally formed five groups: (1) Spaniards [denoted SPAIN], (2) European Union but not Spaniard, North America, and Oceania [denoted EUNAO]; (3) Europe not EU, Eastern Asia, and South-Eastern Asia [denoted EAST]; (4) Latin America and the Caribbean [denoted LAC]; and (5) Northern Africa, Central Asia, Western Asia, and Sub-Saharan Africa [denoted REST]. The classification of countries or areas into geographical regions follows the M49 standard of the United Nations. Following a worst-case logic, the 48 individuals with non-available value have been imputed with value "Northern Africa, Central Asia, Western Asia, and Sub-Saharan Africa", the national group with the worst prospect following Polavieja et al. (2023).

Level of education (qualitative, ordered). This variable measures the level of education following ISCED 2011 but making two groups to avoid small sample sizes: levels 0 and 1, and levels 7 and 8.

Disability (qualitative). It is a binary variable that collects if the individual has a recognized degree of disability greater than 33 %, which is the threshold established in Spain for entitlements of persons with disability.

Local labour market (qualitative). The current datasets do not offer information on the municipality of residence of each individual updated for each quarter. For the treatment group, we take the municipality of residence of the participant from the GIA dataset. For the control group, we collect the information from the DEMANDS dataset, which observes it at different dates. We employ the classification of municipalities into local labour markets (LLM) developed by Martínez-Bernabéu and Casado-Díaz (2022). Although we departed from the 27 local labour markets detected with this method, we had to collapse some categories to avoid small sample sizes. These collapses were made according to geographical distance and similarity in terms of productive structure (proportion of big sectors). Thereby, we ended up with 14 categories: LLM 08305 (Vilafranca del Penedès), LLM 08019 – Barcelona, LLM 08113 – Manresa, LLM

08298 – Vic, LLM 43161 – Valls, LLM 43148 – Tarragona, LLM 17079 – Girona, Rest of Barcelona (LLM 08102 – Avinyó and LLM 08022 – Berga), Ebre (LLM 43155 – Tortosa and LLM 43093 – Móra d'Ebre), Coast of Girona (LLM 17117 – Palafrugell, LLM 17160 – Sant Feliu de Guíxols and LLM 17023 – Blanes), Interior of Girona (LLM 17066 – Figueres, LLM 17147 – Ripoll and LLM 17114 – Olot), Lleida and Mollerusa (LLM 25120 – Lleida and LLM 25137 – Mollerusa), Interior of Lleida (LLM 25217 – Tàrrega, LLM 25072 – Cervera, LLM 25234 – Tremp, LLM 25207 – Solsona and LLM 25203 – La Seu d'Urgell), and Outside Catalonia (LLM 22061 – Binéfar, LLM 44013 – Alcañiz, LLM 22112 – Fraga).

Job search area (qualitative). The available dataset allows us to distinguish seven categories, with one of them without a specific geographic area, which makes impossible to think this variable as ordinal. The categories are municipality, county, province, region (Catalonia), Spain, bigger than Spain, and other areas.

Days with benefit in the year before treatment (quantitative). This variable gathers the number of days in which the individual received at least one of the benefits managed by the public employment service of Spain (SEPE). This includes contributory benefits (like the unemployment insurance) and non-contributory benefits (like the unemployment subsidy).

Unit component (quantitative). This variable is a proxy for unobserved variables that could act as confounders, like ability or personality. These must be variables whose value was time-invariant and whose effect was time-invariant.

Time component (quantitative). This variable is a proxy for unobserved variables that could act as confounders, like the presence of a recession of the business cycle.

A.4. Summary statistics of our sample

Table A.4.1 – Summary statistics of treatments by confounders or proxies of confounders (qualitative variables).

		n	*Percentage*
Sex	Man	15,246	41.2
	Woman	21,791	58.8
National group	EAST	459	1.2
	SPAIN	31,750	85.7
	EUNAO	1,171	3.2
	LAC	2,128	5.7
	REST	1,529	4.1
Level of education	ISCED 0 and 1	976	2.6
	ISCED 2	15,951	43.1
	ISCED 3	10,116	27.3
	ISCED 5	4,866	13.1
	ISCED 6	2,183	5.9
	ISCED 7 and 8	2,945	8.0
Disability	No	33,774	91.2
	Yes	3,263	8.8
Local labour market	Barcelona	24,140	65.2
	Manresa	1,024	2.8
	Vic	716	1.9
	Vilafranca del Penedès	1,140	3.1
	Girona	1,204	3.3
	Tarragona	3,071	8.3
	Valls	556	1.5
	Coast of Girona	1,093	3.0
	Ebre	1,602	4.3
	Interior of Girona	214	0.6
	Interior of Lleida	407	1.1
	Lleida and Mollerusa	1,377	3.7
	Outside Catalonia	32	0.1
	Rest of Barcelona	461	1.2
Job search area	Municipality	493	1.3
	County	22,207	60.0
	Province	7,060	19.1
	Catalonia	572	1.5
	Spain	276	0.7
	Bigger than Spain	314	0.8
	Other areas	6,115	16.5

Note: Summary statistics for the analysis sample of treated units used in our study, i.e. those that have common support and passed the requirements indicated in Section 3.1. See the previous part for more information on the variables.

Table A.4.2. Summary statistics of treatments by confounders or proxies of confounders (quantitative variables).

	n	*Percentage*	*Mean*	*Min*	*Q1*	*Med.*	*Q3*	*Max*
Age			38.4	14	29	39	48	72

Note: Analysis sample of treated units used in our study.

Table A.4.3 – Summary statistics of treatments by cohort.

	n	*Percentage*
8 (2016T4)	3,445	9.3
9 (2017T1)	10,156	27.4
10 (2017T2)	5,318	14.4
11 (2017T3)	876	2.4
12 (2017T4)	4,114	11.1
13 (2018T1)	8,105	21.9
14 (2018T2)	4,374	11.8
15 (2018T3)	649	1.8
Total	37,937	100

Note: Analysis sample of treated units used in our study.

Table A.4.4 – Summary statistics of treatments by duration, specificity, and internships.

	Mean	*Min*	*Q1*	*Med.*	*Q3*	*Max*
Duration (hours)	568	150	390	590	800	920
Specificity (# linked occupations)	4.11	1	2	5	5	8
Internships (proportion of hours)	13.2	0	9.3	15	15	40

Note: Analysis sample of treated units used in our study.

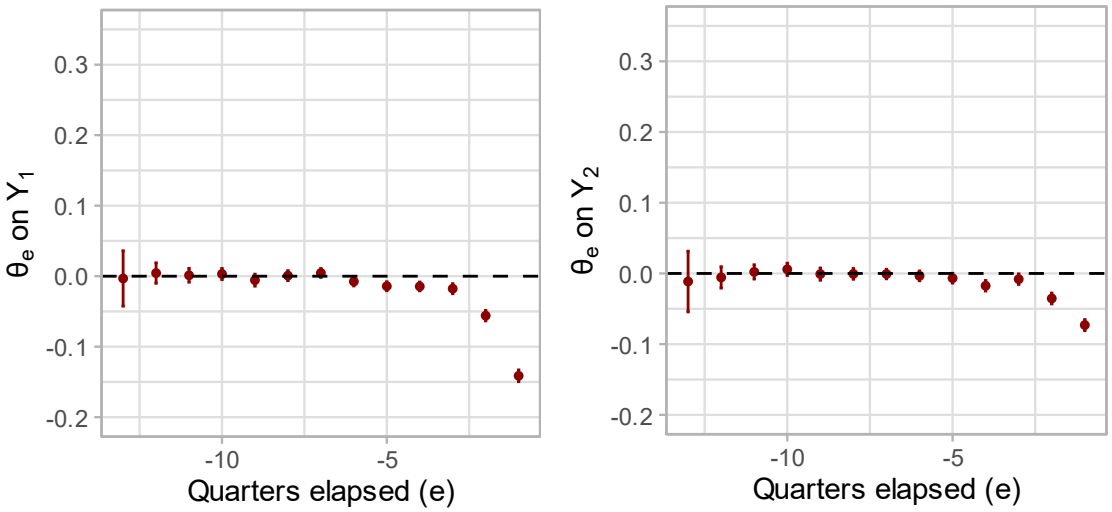
Appendix B. On the anticipation parameter

Following the rationale of Callaway (2021), we have plotted pre-treatment outcome differences with varying base periods to select the anticipation value (δ). The base period is the time period at which the unit fixed effect is estimated, i.e. it is the moment at which the permanent differences between units is estimated. If there is no anticipation, researchers usually use the last period before treatment as base. Event-study plots employ a universal base period to evaluate the plausibility of the bias stability (parallel trends) assumption at pre-treatment periods in a sort of placebo outcome analysis. For pre-treatment periods, the varying base period configuration changes the base period at each period of estimation of outcome differences to the previous period. This helps us to examinate how stable are these differences across periods.

According to Figure B1, although we appreciate minor differences from one period to the next in quarters -7 to -3, we see that the big changes happen at quarters -2 and -1. This makes us think that δ should be fixed at 2. This is the conventional approach to address the Ashenfelter's dip in labour economics (Malani & Reif, 2015). This pattern is repeated across cohorts (g)[59], which induces us to think that the Ashenfelter's dip is not product of a random economic shock. We interpret it as an interval of search (frictional) unemployment in which the search is directed towards finding a proper training course as a substitute for job search.

[59] These cohort-specific plots are available upon request.

Figure B1 –Estimates of ATT_e of the participation in FOAP on Y_1 and Y_2 for $e < 0$ with varying base periods.

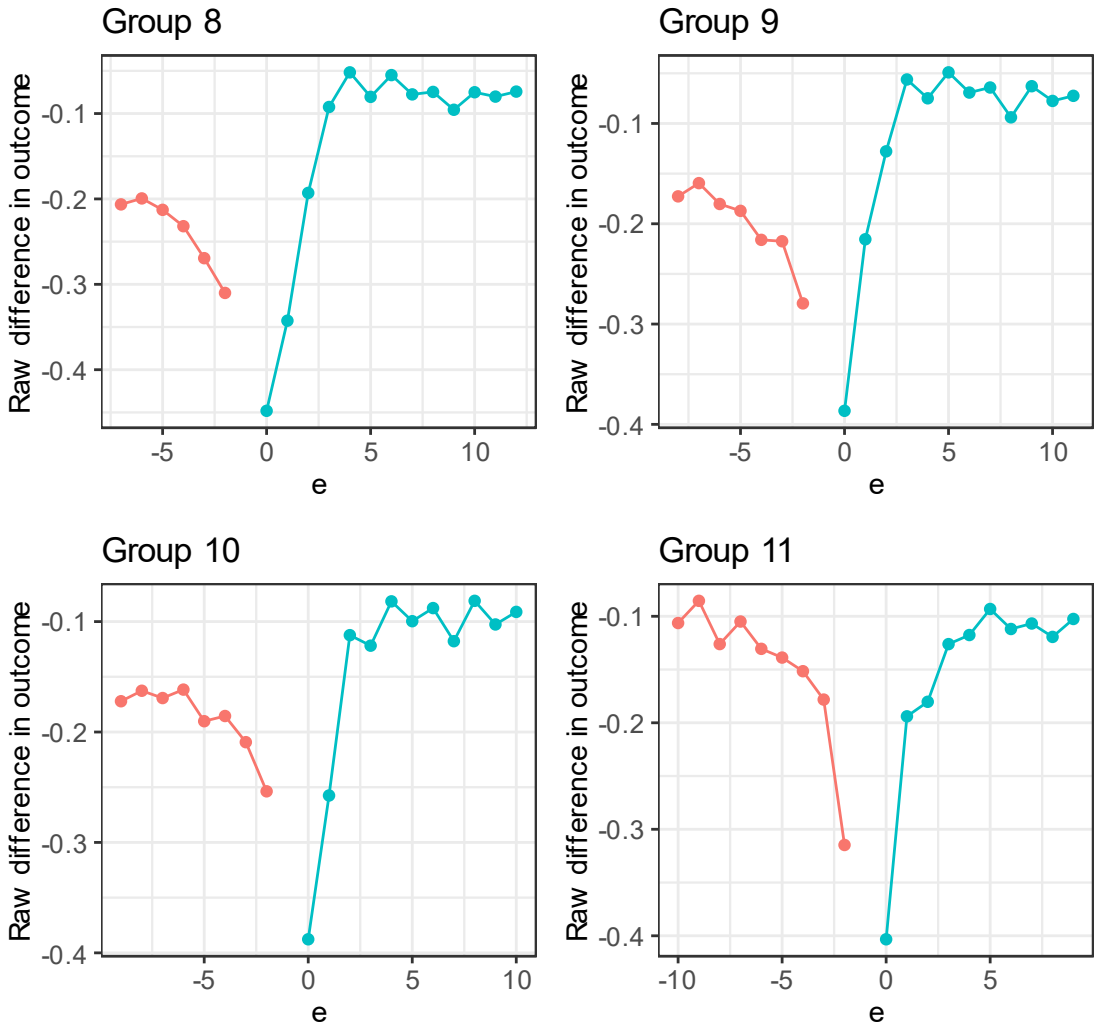


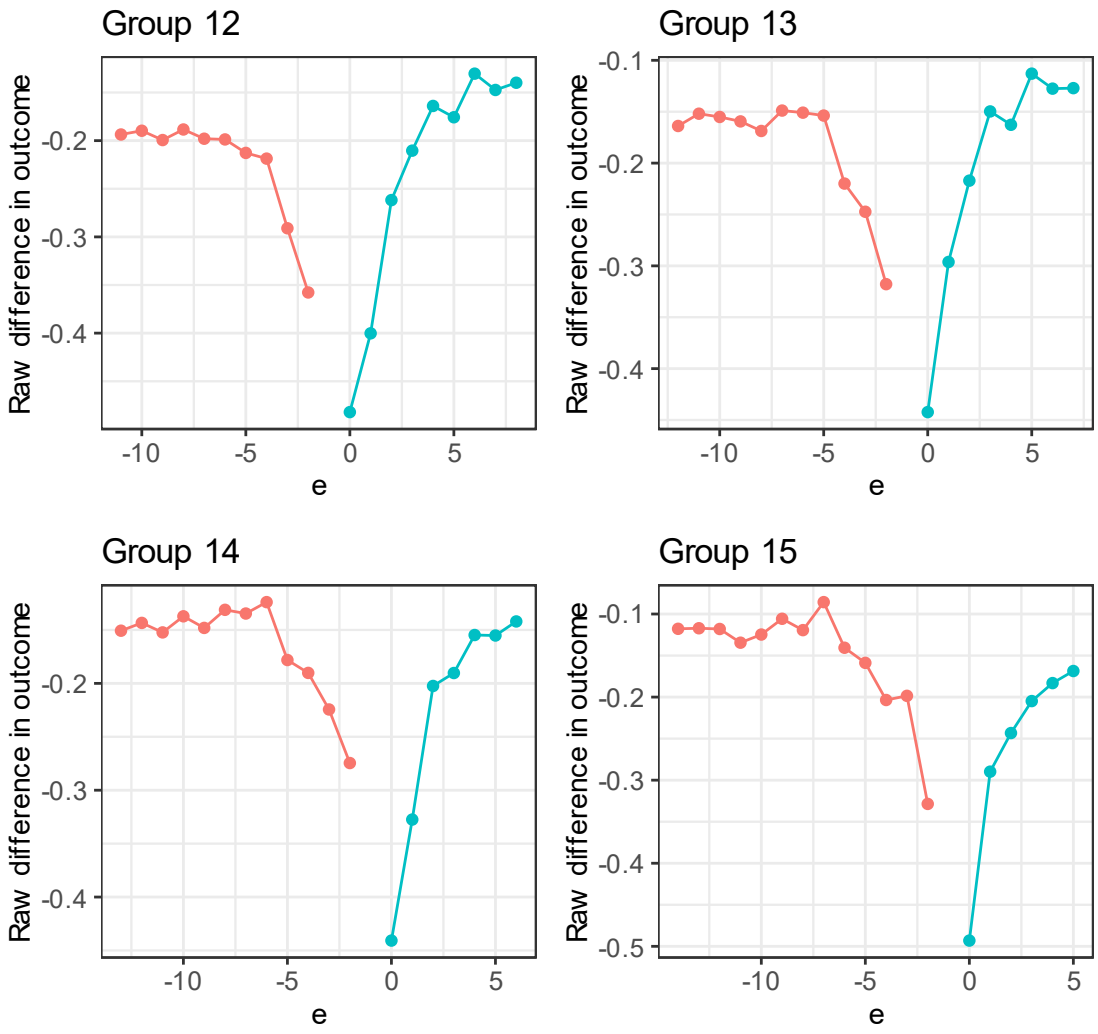
Note: It is shown the point estimate and the 95 % confidence interval.

Appendix C. Additional tables and figures

C.1. Raw differences between treatment and control group

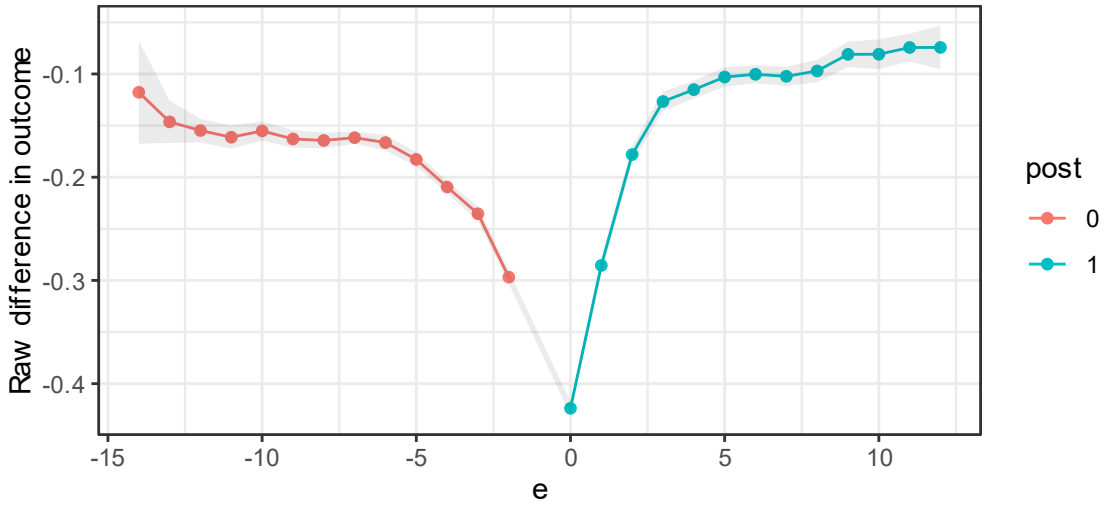
Figure C.1.1. Raw differences in Y_1 between treatment and control group for each g and e.





Note: Pre-treatment outcomes in red, post-treatment outcomes in blue.

Figure C.1.2. Raw differences in Y_1 between treatment and control groups for each e.



Note: This plot shows an aggregated parameter across $G = g$.

C.2. Estimates of θ_e for both identification strategies

Table C.2.1 – Estimates of θ_e on employment (Y_1) and stable employment (Y_2) by identification strategy.

	CDID		*CLDV*	
	Employment	*Stable employment*	*Employment*	*Stable employment*
$\theta_{e=0}$	-0.166***	-0.089***	-0.137***	-0.098***
	(0.004)	(0.003)	(0.002)	(0.002)
$\theta_{e=1}$	-0.004	0.040***	-0.052***	-0.003*
	(0.004)	(0.004)	(0.003)	(0.002)
$\theta_{e=2}$	0.109***	0.108***	0.035***	0.045***
	(0.005)	(0.004)	(0.003)	(0.003)
$\theta_{e=3}$	0.175***	0.135***	0.078***	0.057***
	(0.005)	(0.004)	(0.003)	(0.003)
$\theta_{e=4}$	0.215***	0.140***	0.062***	0.036***
	(0.006)	(0.005)	(0.004)	(0.003)
$\theta_{e=5}$	0.245***	0.152***	0.051***	0.03***
	(0.007)	(0.007)	(0.004)	(0.003)
$\theta_{e=6}$	0.256***	0.164***	0.052***	0.03***
	(0.007)	(0.007)	(0.004)	(0.004)
$\theta_{e=7}$	0.259***	0.179***	0.045***	0.032***
	(0.006)	(0.006)	(0.004)	(0.004)
$\theta_{e=8}$	0.277***	0.192***	0.03***	0.036***
	(0.008)	(0.007)	(0.005)	(0.004)
$\theta_{e=9}$	0.286***	0.198***	0.029***	0.044***
	(0.008)	(0.009)	(0.005)	(0.005)
$\theta_{e=10}$	0.290***	0.197***	0.039***	0.041***
	(0.008)	(0.009)	(0.005)	(0.005)
$\theta_{e=11}$	0.301***	0.208***	0.045***	0.046***
	(0.009)	(0.009)	(0.006)	(0.005)
$\theta_{e=12}$	0.342***	0.234***	0.035***	0.046***
	(0.012)	(0.012)	(-0.137)	(0.009)

Note: The column on CDID uses $\delta = 2$, while the column on CLDV uses $\delta = 0$. The rest of results with different δ are available upon request.

C.3. Estimates of $\theta_{g,e=5}$ for the CLDV identification strategy

Table C.3.1. – Estimates of $\theta_{g,e=5}$ on Y_1 and Y_2 by δ.

	$\delta = 2$		$\delta = 0$	
	Employment	*Stable employment*	*Employment*	*Stable employment*
$\theta_{g=8,e=5}$ (2016T4)	-0.07***	-0.044***	0.01*	-0.005*
	(-0.005)	(0.004)	(0.005)	(0.004)
$\theta_{g=9,e=5}$ (2017T1)	-0.021***	-0.01***	0.045***	0.023***
	(0.003)	(0.003)	(0.003)	(0.003)
$\theta_{g=10,e=5}$ (2017T2)	-0.061***	-0.027***	0.026***	0.015***
	(0.005)	(0.004)	(0.005)	(0.004)
$\theta_{g=11,e=5}$ (2017T3)	-0.087***	-0.047***	0.011*	0.018***
	(0.012)	(0.008)	(0.011)	(0.009)
$\theta_{g=12,e=5}$ (2017T4)	-0.13***	-0.058***	-0.049***	-0.013***
	(0.005)	(0.004)	(0.005)	(0.004)
$\theta_{g=13,e=5}$(2018T1)	-0.094***	-0.026***	-0.017***	0.012***
	(0.005)	(0.004)	(0.005)	(0.004)
$\theta_{g=14,e=5}$(2018T2)	-0.119***	-0.042***	-0.012**	0.011***
	(0.006)	(0.005)	(0.006)	(0.005)
$\theta_{g=15,e=5}$(2018T3)	-0.16***	-0.057***	-0.005*	0.007*
	(0.013)	(0.01)	(0.013)	(-0.011)

Note: Own elaboration. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate and the standard error in parenthesis.

Table C.4.1 – Additional estimates of $ACR^{(k)}_{d;j,e}$ for Y_1 at $e = 3$ in function of the identification strategy and the anticipation parameter

CDID and $\delta = 0$			
Δ_1	-0.033***	-0.021***	0.094***
(250-150; 2-1; 10-0)	(0.007)	(0.006)	(0.014)
Δ_2	-0.051***	-0.020***	0.038***
(500-250; 4-2; 20-10)	(0.009)	(0.006)	(0.007)
Δ_3	-0.008	0.008	-0.018*
(750-500; 6-4; 30-20)	(0.007)	(0.006)	(0.010)
Δ_4	0.019*	0.036**	-0.074***
(920-750; 8-6; 40-30)	(0.011)	(0.014)	(0.018)
Passed the placebo test	Yes	Yes	No

Chapter III.

Evaluating two types of profiling models for the public employment service of Catalonia[60]

1. Introduction

Public employment services (PES) usually pursue the reduction of long-term unemployment (LTU) as one of their objectives. From an individual perspective, researchers have estimated that a prolonged unemployment episode affects economic outcomes like the employment probability (Filomena, 2023) and health outcomes like mental health (Picchio & Ubaldi, 2022). From a social perspective, longer unemployment episodes entail greater costs in PLMPs and other welfare benefits, not to mention the risks of social exclusion and vulnerability faced by the long-term unemployed. Countries in Southern Europe like Spain or Italy multiply by two the average of the EU for the annual long-term unemployment rate (Eurostat, 2024c). The consequence has been that Spain ranked at the second position in the European classification of expenditure in passive measures, with the 1.52 % of its GDP in 2019. Simultaneously, these countries have a relatively low spending in ALMPs like job search services (DG EMPL, 2024). This combination of phenomena requires ensuring an efficient distribution of jobseekers to interventions.

One way to improve the efficiency of the programs is to detect which persons have a higher probability of LTU through profiling models. Once they are identified, PES may provide them a faster, more intense or different intervention. Thereby, prediction turns a key task for these agencies to decrease the duration of unemployment. Machine learning models have been especially used for this goal in different policy fields (Kleinberg et al., 2015), and some PES have already started to use these modern approaches to improve their services (Körtner & Bonoli, 2023). Researchers or practitioners have already analysed administrative data in countries like Portugal (de Troya et al., 2018), France (Gallagher & Griffin,

[60] This chapter is based on the working paper titled "From rules to forests: rule-based versus statistical models for jobseeker profiling", which is coauthored with Christoph Kern (LMU).

2023), Belgium (Desiere & Struyven, 2021), or New Zealand (Desiere et al., 2019) to develop these modern statistical profiling systems.

Nonetheless, it is not obvious that statistical models must outperform caseworker or rule-based systems, which are the types of jobseeker profiling already implemented in some countries (Loxha & Morgandi, 2014). On one side, statistical models are usually fed by variables easily collected by public administrations. On the other side, caseworker and rule-based systems may use more difficult-to-measure factors to predict LTU by leveraging the expertise of counsellors. The problem is that PES rarely document at least publicly these procedures, which makes difficult to compare the performance of each approach. As far as we know, there have been only two studies contrasting the predictive performance of different types of jobseeker profiling and both deal with two countries from Continental Europe (van den Berg et al., 2024; Desiere & Struyven, 2021).

With this chapter we contribute to the literature in three directions. Firstly, we contrast the performance of the current rule-based profiling model of Catalonia versus our statistical models by using a new dataset formed by bringing together many administrative sources. Secondly, we expand the set of metrics considered to evaluate the predictive performance by considering both the discrimination ability and the calibration of the models. The calibration approach emphasizes the importance of giving accurate numeric estimates on the probability of the event, since this informs the person on the uncertainty of the diagnostics. For instance, a jobseeker classified in the high-risk group might reconsider her decision to move to another region if she knew that her LTU probability is of 55 %. Thirdly, we develop and validate the first statistical profiling models for Catalonia and the first machine learning profiling models for Spain. We find that our statistical models remarkably outperform the current rule-based system by attaining up to a ROC-AUC of 0.735 as compared to the 0.593 found for the rule-based procedure. We highlight the gains of the statistical models with additional discrimination metrics, but the benefits of the machine learning approach are especially salient regarding calibration. Still, we discuss the importance of evaluating the models with regards to the specific profiling setup through techniques like stress tests or model interpretation statistics.

The next section selectively reviews the literature on jobseeker profiling by defining the main concepts and indicating the main references. We continue with

a description of the data and the methods employed to develop and validate our models. The following section presents the findings of the analysis by indicating the performance metrics of the models. We pay special attention to the selection of models that might seem equivalent according to these statistics. To end up, some conclusions and avenues for further research are offered.

2. The Catalan profiling model in context

2.1. Jobseeker profiling

In relation to labour market policies, profiling models are employed to classify jobseekers via classes (e.g., low or high probability of LTU) or scores (e.g., the probability of LTU) (Körtner & Bonoli, 2023)[61]. These classifications are typically used to describe more briefly the set of jobseekers, to distribute resources across people, or as an intervention that updates the information of the individual on its employment prospect (Loxha & Morgandi, 2014; Harmon et al., 2021)[62]. In short, we may summarize these three uses as: data reduction, allocation, and active diagnostics. Throughout this research, we focus on the allocation use because it is the one closest to the task of creating target groups. In the allocation problem, jobseekers often face a sequence of interventions in which they may participate at different times[63]. Therefore, we can distinguish an intra-sequence of actions (for the same individual, ordering interventions across time) and an inter-sequence of actions (for the same intervention and time, an order of individuals).

Departing from Kuppler et al. (2022), we can frame the allocation problem in PES as an example of data-driven decision making and talk about a two-step

[61] The main point is to *classify*. Classes, groups, segments, subsets, categories, or clusters are usually treated as synonyms and defined as elements of a *finite* ordered or unordered set. Scores are elements of an *infinite* ordered set of numbers, like $\{x \in \mathrm{R} \mid 0 \leq x \leq 1\}$.

[62] The European Union (2016) in the General Data Protection Regulation defines profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements". Although it circumscribes profiling to *automated* processes, caseworkers may also profile without the use of computers.

[63] Profiling may also be used to allocate job offers, which is called sometimes labour market matching (Körtner & Bonoli, 2023; Loxha & Morgandi, 2014). We focus on the distribution of jobseekers into interventions. In ALMP, some interventions like employment subsidies are very similar to a job offer.

procedure. The first stage is the *decision* step, in which the PES establishes variables relevant for the allocation and a mathematical function to map the value of the individual in these variables (the input) on the interventions of the agency (the output). This function or allocation principle takes the value and allocate resources, so it can be judged based on different distributive justice theories (Kuppler et al., 2022). In most cases, this set of relevant variables reduces to variable about the jobseeker in the future (not at the time of the allocation t), i.e. the employment situation of the individual at time $t + e$. Since we have to approximate a value in the future, we need a predictive model. In other cases, the decision-relevant variables deal with the situation of the worker at the decision time, so we require a descriptive model. This descriptive model might use Boolean expressions of manifest variables (e.g. man and older than 45 to group A) or estimate the membership of an individual in a latent group (Fernández et al., 2018). The second stage is then the *profiling* step. If it uses a predictive model, which is by far the most common, its evaluation will compare the true and the predicted value. If it employs a descriptive model, the objectives of the grouping can vary, therefore the method of evaluation will also differ (Hennig, 2015). In any case, we borrow from Kuppler et al. (2022) the concept of *allocation system* to integrate in the same workflow the profiling model and the decision model.

Research on jobseeker profiling has mainly interested on the types of profiling models and their predictive performance. Literature reviews on jobseeker profiling can be found in Duell & de Moraes (2023), Desiere et al. (2019), or Barnes et al. (2015). For this reason, here we concentrate on two research lines that have been addressed in our work: the predictive performance of these models according to how much human discretion they incorporate and the implementation of profiling models in PES of our context.

Profiling models are usually classified into three types in function of how much human discretion they incorporate: caseworker-based, rule-based, or statistical profiling (Desiere et al., 2019; Rebollo-Sanz, 2018)[64]. In caseworker procedures, each counsellor may have its own function to allocate individuals into groups, enjoying a high level of autonomy and discretion. This means that each caseworker might consider different variables as relevant, with the allowance to even vary this set of relevant criteria among jobseekers. Moreover, the weight attached to each

[64] The same differentiation of three types may be applied to decision models.

of these covariates might also vary from counsellor to counsellor. On the contrary, rule-based and statistical systems establish a unique function for all counsellors and jobseekers. Whereas statistical models determine the importance of each variable through data analysis, rule-based models establish the parameters of the function through the human intuition of policymakers or through any non-statistical technique (Rebollo-Sanz, 2018). Thereby, all types of jobseeker profiling entail human discretion, but rule-based and statistical models limit it to the decision step. In this chapter, we compare the performance of two examples of the two standardized profiling types: the rule-based and the statistical procedures.

To evaluate the performance of a profiling model, researchers often employ discrimination metrics[65]. Rule-based models have been implemented at least in Ireland, Poland, Norway, and United Kingdom (Desiere et al., 2019; Loxha & Morgandi, 2014). We believe that they are even more prevalent, since many ALMP in Europe incorporate eligibility criteria that may be recognized as part of an implicit rule-based system (Cronert, 2022). The problem is that these models have rarely been described in detail and their assessment is even scarcer. As far as we know, Desiere & Struyven (2021) is the unique work analysing the discrimination metrics of a rule-based model. They study a system implemented in Belgium and find that it presents an accuracy of 0.58 for predicting employment after six months, with greater false alarm rate for foreign people. However, they discover that an alternative statistical model presents the same false alarm rate but increases the accuracy of the predictions. Even so, it should be noted that they study a very straightforward rule-based system. Other PES like the Catalan have implemented more sophisticated versions of rule-based models[66].

The use and evaluation of statistical profiling models is wider. According to Desiere et al. (2019), they have been deployed in the PES of Austria, Belgium, Denmark, Ireland, and Netherlands (Desiere et al., 2019). Their analyses indicate that they attain an accuracy that goes from 0.61 to 0.85. Moreover, researchers have lately developed new models for countries like Germany, with a ROC-AUC of 0.75-0.77 (Bach et al., 2023; Kunaschk & Lang, 2022); Finland, achieving a ROC-AUC of 0.8 (Viljanen & Pahikkala, 2020), and Slovakia, with an accuracy equal to 0.918 (Gabrikova et al., 2023). Some studies have even compared the

[65] In biostatistics, discrimination is the ability to separate units that will and will not suffer the event of interest (Austin & Steyerberg, 2012).

[66] Appendix A includes a figure depicting one of the rule-based models analysed in this chapter.

performance of caseworker-based versus statistical models and have found that the last ones have much more sensitivity (van den Berg et al, 2024; Arni & Schiprowski, 2015). In the Appendix B, we offer a systematic review on the performance metrics found in the literature and we show that the ROC-AUC are always between 0.7 and 0.8.

In the context of Spain, there is a low implementation of profiling models and statistical versions have only been developed by Felgueroso et al. (2018) and Molina et al. (2023)[67]. Although both studies used generalized linear models, they employed different predictors. Felgueroso et al. (2018) estimated models with predictors widely used in labour market research and suggested that age was one of the best predictors of LTU[68]. Casanova et al. (2021) reported the use of a related profiling model in the implementation of a private-provided ALMP in Catalonia. Molina et al. (2023) focused on three alternative sets of predictors related to the personality, the personal networks, and the employment expectancies of jobseekers. Consistent with the results of van den Berge et al. (2024), this study found that expectancies were a great predictor of LTU. Our chapter tries to integrate conventional and usually unobserved covariates by introducing lagged outcomes (Caliendo et al., 2017).

[67] The central PES (*Servicio Público de Empleo Estatal, SEPE*) has created an application called Send@ more similar to a targeting system than to a profiling system according to the definitions of Körtner & Bonoli (2023). In short, profiling models estimate a potential outcome after no intervention ($\Pr[Y(0) = y|X]$) and targeting models estimate a vector made up by one potential outcome after each intervention $\boldsymbol{d}$ ($\mathbf{r} = \Pr[Y(d_1) = y|\boldsymbol{X}], \Pr[Y(d_2) = y|\boldsymbol{X}], ..., \Pr[Y(d_K) = y|\boldsymbol{X}]$). Following Muñiz (2021), a caseworker introduces the values that the jobseeker has in certain covariates ($\boldsymbol{X_{it}} = (age_{it}, education_{it}, search_area_{it}, languages_{it})$) and then the application estimates the potential outcome for each intervention fixing $\boldsymbol{X} = \boldsymbol{x}$. Then, Send@ outputs two vectors sorted from more to less estimated probability. First, a vector (list) on the estimated probabilities in function of each intervention ($\mathbf{r}$). Second, a vector on the estimated probabilities in function of each target occupation $\boldsymbol{o}$, e.g. $\widehat{\Pr}[Y(o_1) = y|\boldsymbol{X} = \boldsymbol{x}]$. This is where this application moves further away from a classical targeting system. In our opinion, Send@ may confuse their users on the interpretation of these quantities by understanding them as intervention effects instead of mere predictions.

[68] All their covariates were introduced in the models as categorical, so it is challenging to assess which variable has greater prediction ability. The reasons are that not all covariates have the same degrees of freedom and only "average marginal effects" (average risk differences) are presented.

2.2. The Catalan profiling model

The PES of Catalonia (called SOC) already uses an allocation system. The design of the current profiling model states that their objectives are to increase the take-up, the effectiveness, and the efficiency of the services (SOC, 2016). As we introduced before, profiling models may be used to determine *where* to go (which interventions, the intra-sequence) and to determine in which position of the *queue* (when to be treated, the inter-sequence). Table 1 summarizes the main features of the Catalan allocation system, with the profiling and the decision models depicted with more detail in Appendix A.

On the profiling side, a rule-based procedure classifies individuals for the first intervention by assigning them to a certain group through function Q-G and by providing them a certain score through function Q-S. The rule-based procedure is different in the first and the second action, but we concentrate on the model for the first intervention since this profiling provides a label to the individual that will be continuously attached to her record. In the next interventions, the profiling may be done in a caseworker-based way, but her Q profiles will still be updated to monitor the progress of the jobseeker. Caseworker profiling is not documented, so it cannot be assessed. Thereby, Q models are considered the main profiling device (SOC, 2016).

On the decision side, the relevant criteria are (a) for the sequence of interventions, a *vector* of employment mediators like the target occupations or the level of education, and (b) for the sequence of individuals, a scalar of the employment probability. Once the profiling is done, another rule-based procedure indicates the first and second interventions for each unemployment episode of each individual who wishes to participate in ALMPs[69]. For these interventions, it locates the jobseeker in certain actions of job search assistance. For the remaining interventions (if necessary), the profiles of individuals created during this first prediction step are offered to the caseworker to assist the decision on future interventions (SOC, 2016).

[69] Sometimes the rule-based model only determines the first intervention (see Appendix A).

Table 1 – Scheme of the allocation system currently implemented by SOC for intervention 1

	Decision-relevant criteria	*Profiling function*	*Decision function*
Intra-sequence of interventions	Mediators of employment	Q-G	D_1
Inter-sequence of interventions	Pr [*employment*]	Q-S	Caseworker-based

Note: The rows collect the dimension of the allocation system (intra or inter) and the columns the steps of the system. For the second intervention, C substitutes Q-G and D_2 substitutes D_1. Functions Q-G, Q-S, D_1, D_2, and C are formally defined in the Appendix.

Q functions may be disentangled in inputs, processing, and outputs. The *inputs* of these models are variables on labour markets from administrative sources and other variables gathered via questionnaires to jobseekers. The administrative data mainly collect information about the economic context at present, like the unemployment rates by occupation and industry. The questionnaire data includes information on skills, work experience, and additional credentials[70]. The *processing* of the data is done by different functions, all of them with parameters set according to rules (human choices) and not to statistical techniques. Q-G may be understood as a step function or a set of Boolean expressions, which can be easier represented through a decision tree. Q-S is just a linear and additive function with parameters assigned to each qualitative variable. Appendix A includes the mathematical expressions of both functions. The *outputs* are again two: $G_{it} \in \{\text{A1}, \text{A2}, \text{A3}, \text{A4}, \text{B1}, \text{B2}, \text{B3}, \text{C1}, \text{C2}, \text{D}, \text{Z1}, \text{Z2}, \text{Z3}, \text{E}, \text{R6}\}$ for Q-G, which is informally called the subgroup of an individual[71], and $S_{ito} \in [0,139]$ for Q-S. G_{it} is then conceptualized as a latent variable or simply a combination of employment mediators and S_{ito} as the employment probability but in a different scale. Despite that, we will argue that we may transform them easily to interpret both functions as predictive proxies of the probability of LTU. On one side, S_{ito} is defined for each individual i, time t, and occupation o. On the other side, G_{it} is defined for

[70] Appendix D includes an exhaustive list of the variables included in Q models.

[71] In Spanish, *subcolectivo*.

each individual i and time t. However, to reduce the notational burden, we will remove the subindexes when they are not strictly necessary to follow the thread.

When is this allocation system applied to jobseekers? The design document states that S and G may be calculated at most once a month (SOC, 2016). Some reasons that justify a new calculus of the profile are starting an intermediation claim[72], varying such claim, or finishing an ALMP. However, the implementation evaluation of Everis (2017) on this profiling model reported that caseworkers manually altered the group (G) in one out of five cases. A possible explanation is that 45 % of these civil servants considered that the efficacy of this profiling was low or moderate.

3. Methods

3.1. Data

We have used administrative microdata granted by the Catalan PES. Specifically, we merged four datasets to develop our models: intermediation claims (SICAS), employment contracts (Contrat@), participations in ALMPs (Galileu), and participations in PLMPs funded by the central government. To build the final dataset, different steps have been followed. Remind that the goal is a panel dataset whose units of analysis (rows) are unemployment episodes with a variable (column) gathering the duration of the episode and covariates (rest of columns) measured at the starting day of such unemployment episode.

Firstly, the data provider got one simple random sample of 25,000 jobseekers for each focal year (2017, 2018, 2019, and 2022). It was a sample from the population of persons registered at least once as unemployed in the focal year. Secondly, for the people in the sample, the data provider extracted the variables requested for each sampled jobseeker for the time window of interest, i.e. [2015, 2023]. The data provider gave us the raw files and we continued with the data wrangling. Thirdly, we constructed a dataset of labour market episodes and a dataset of public policy episodes. We define an episode as the participation of individual i in a certain event during a closed interval of time that begins at day t and finishes at day $t' \geq t$. The events of interest for the first episodes are participations in the labour market, while those of interest for the second episodes are participations in

[72] In Spanish, *demanda de empleo*.

ALMPs or PLMPs. Our definition of labour market episodes makes them time intervals that may not overlap over time, whereas policy episodes can do it.

Let us review in more detail each type of episode. On one side, there are three subtypes of labour market episodes: employment, unemployment, and inactivity. A new labour contract determines the starting of a new employment episode. Certain types of intermediation claim initiate a new unemployment episode, whereas others start an inactivity episode. The justification is that each intermediation claim includes the reason of the registration, suspension, or removal of the intermediation from the internal system of the PES[73]. For instance, changes in the intermediation claim motivated by temporary working inability, family care, or prison entry are considered as inactivity reasons. Therefore, a new unemployment or inactivity episode is registered by the presence or sequence of intermediation claims coded with such reason. On the other side, public policy episodes are differentiated in four subtypes: job search interventions, adult training, employment subsidy, or receiving a benefit.

Table 2 – Sample sizes and events of interest by year in which the episode started.

Year	*Unemployment episodes*	*LTU episodes*	*Individuals*	*Individuals with at least one LTU ep.*
2017	44,852	9,757 (21.8 %)	31,524	9,757 (30.95 %)
2018	46,548	9,468 (20.3 %)	32,639	9,468 (29.01 %)
2019	47,648	11,618 (24.4 %)	33,656	11,618 (34.52 %)
2020	57,473	18,825 (32.8 %)	37,096	18,825 (50.75 %)
2021	32,985	7,612 (23.1 %)	23,355	7,612 (32.59 %)
2022	34,922	5,803 (16.6 %)	24,952	5,803 (23.26 %)
Total	292,725	63,083 (21.55 %)	85,398	54,781 (64.15 %)

The fourth step was filtering the dataset of labour market episodes to a dataset of unemployment episodes. The remaining information was incorporated as new columns, i.e. as new predictors. For the left-hand side of the regression equations, we choose as outcome the event of long-term unemployment in line with most

[73] Appendix C includes a table that relates each type of intermediation claim with the respective code of unemployment or inactivity.

studies of jobseeker profiling (Körtner & Bonoli, 2023). Respecting the definition of Eurostat (2024b), we considered that there was LTU if the unemployment lasted at least 365 days. Table 2 shows our sample sizes and the prevalence of LTU for each year.

For the right-hand side, the predictors collected in our data may be grouped in the sets presented in Table 3. The exhaustive list of predictors is presented in Appendix D and is greatly influenced by Bach et al. (2023), who conducted a similar study for Germany. There are time-invariant covariates (like sex or nationality), but also time-variant variables whose values are calculated for each episode. Inside this last group of time-variant covariates, we find those columns that gather the histories of labour market and policy episodes. Appendix E offers summary statistics on the socio-demographic features of our sample.

Table 3 – Groups of predictors

Group	*Number of predictors*	*Predictors (examples)*
Employment	18	Days since last employment, days since last full-time employment, occupation of last employment…
Unemployment	5	Total duration of unemployment episodes, number of unemployment episodes, days since last unemployment episode…
Inactivity	3	Total duration of inactivity episodes, number of inactivity episodes, mean duration of inactivity episodes.
PLMP	5	Started the unemployment episode during a benefit interval, number of benefit episodes completed, total duration of benefit episodes…
ALMP	9	Total duration of job search assistance episodes, total duration of adult training episodes, total duration of employment subsidy participations…
Socio-demographics	37	Sex, nationality, age, field of education…

Note: Appendix D shows the complete list of predictors.

These data were used to develop our prediction models, but we also need data on the output of the current profiling models to make the performance comparison. We have data for all the calculations of Q-G and Q-S carried out during the temporal window of interest. Q-S outputs one S_{ito} for each occupation of interest, so the unit of analysis is different. To place both profiling models with the same unit of analysis, we generated S_{it} as the average of profiling scores output by Q-S for all target occupations[74].

3.2. Analysis

3.2.1. Model development

We have estimated four models to develop our jobseeker profiling, including both classic regressions and tree-based machine learning algorithms: unpenalized logistic regression, penalized logistic regression (Friedman et al., 2010), random forest (Breiman, 2001b), and gradient boosting machine (Chen & Guestrin, 2016). The review by Desiere et al. (2019) concludes that unpenalized logistic regression is the most used technique in our field. We utilize this model as baseline and smooth its problems by additionally estimating its penalized version with an elastic net. The key point is that these conventional models maintain a linear and additive functional form. The advantage is that it makes them easy to interpret, but the disadvantage is that this strict specification is an assumption with a weak justification. Machine learning algorithms allow greater flexibility through non-linearities and interactions at the cost of a lower interpretability. Thereby, the decision on which model to select is not obvious.

Due to our large sample size, we have employed the dataset partition strategy that is common among machine learning practitioners (Kuhn and Johnson, 2019). This involves dividing our dataset in the subsets of training, evaluation, and test. The main reasons to use this technique are avoiding overfitting and assessing the models in a realistic way (Kuhn and Johnson, 2019). The partition is done through two splits. In the first split, we include the unemployment episodes started from 2017 to 2020 in a training plus evaluation subset and apart the unemployment episodes of 2022 in the so-called test subset (Bach et al., 2023). In the second split, we get a stratified random sample to form the training subset (80 % of

[74] The system collects until three occupations of interest.

observations) and the evaluation subset (the remaining 20% of observations). The idea of this last sampling strategy is to ensure a sufficient prevalence of events in each sample, so we use the outcome as the stratifying variable. The training subset is employed both to select the tuning parameters (if there are) and to estimate the regression coefficients. Specifically, tuning parameters are chosen by optimizing the ROC-AUC via temporal cross-validation (Hyndman & Athanasopoulos, 2018)[75]. The evaluation subset is used to pick the probability threshold that will assign a unit prediction to a certain class (e.g., high risk of LTU or low risk of LTU). Lastly, test set is used to evaluate the final models by estimating the performance metrics on this dataset.

To compare the current rule-based model with our statistical models, we need to reduce the test subset to a *restrictive* test subset. We will start by getting the unemployment episodes that were profiled with any Q model in 2022 and then predicting the class or score that they had taken according to our statistical models. This dataset will require two additional filters to attain a fair and useful comparison of profiling models. Let us see the two additional restrictions.

The first restriction allows a proper estimation of performance metrics for a profiling model that has been already deployed in a PES. Performance statistics compare the predicted untreated potential outcome with the factual untreated potential outcome. The problem is that the interventions allocated with the profiling may affect our outcome variable. This means that, if the jobseeker was assigned to a program, we do not observe the untreated potential outcome, but the *treated* potential outcome. For instance, imagine an extreme case of predictive profiling with 100 % of sensitivity and certain ALMP with 100 p.p. of average treatment effect on the treated assigned to those classified as high risk of LTU. If we were to estimate the predictive performance of the profiling model, we would suffer a "blessed curse". A naïve estimation of the sensitivity of the model with the predicted versus the factual (treated) potential outcome would result in an apparent 0 % of sensitivity, since all jobseekers predicted as unemployed after one year were working *thanks to* the ALMP. To avoid this issue, we denote A_{ie} the number of ALMPs in which jobseeker i participated during unemployment episode e. We just need to apply a restriction to the previous dataset: remove those individuals with any participation in ALMPs during the interval of interest

[75] Appendix F includes the grid of candidate values used for the tuning parameters.

$(test \backslash \{A_{ie} > 0\})$[76]. The second restriction limits the test subset to those individuals that are targeted by the Catalan PES. The office does not currently offer its services to jobseekers that do not know neither Catalan nor Spanish, so it redirects them to another public administration. This filter will make the comparison more useful as long as it focuses on the real users of the PES. In any case, we also estimate performance statistics with the non-restricted test subset that includes the individuals excluded in order to offer our findings to a broader audience.

3.2.2. Model validation

For the evaluation of the models, we study two dimensions of predictive performance: discrimination and calibration. Discrimination parameters measure how well a model separates those jobseekers that will and will not experience the event, i.e. LTU in our case. Depending on whether we fix a classification threshold, we may estimate metrics of classification discrimination or ranking discrimination. Calibration parameters concern on how well we approximate the observed proportions of events with the estimated proportions. Since this dimension is typically overlooked in model evaluations, Van Calster et al. (2019) name it the "Achilles heel of predictive analytics" (p. 1). Nonetheless, calibration should be carefully considered by PES, because the uncertainty of an event prediction might make the jobseeker reconsider important decisions like moving or participating in adult training[77]. In a sense, the two dimensions represent different ways of looking at the same topic: as a dichotomous phenomenon (discrimination) or as a latent continuous phenomenon (calibration).

The target parameters on discrimination are the following. On ranking discrimination, we are interested in the area under the receiver operating characteristic curve or c statistic (ROC-AUC) and in the area under the precision-recall curve (PR-AUC). The ROC curve shows the value of sensitivity and false alarm for each possible classification threshold, whereas the PR curve reports the

[76] This would render a fair comparison between profiling models if the currently deployed model did not act as an intervention, i.e. if providing the profile (diagnostic) to the jobseeker did not change her behaviour. Future research might estimate the performance of the current models by simulating the value that an episode would obtain and not by using truly profiled individuals. This is not possible at the moment with the available datasets.

[77] A related approach would be to declare equivocal zones (see Kuhn & Johnson, 2018).

equivalent for precision and sensitivity[78]. The advantage of these statistics is that they do not require taking sides on which classification threshold to set, which is something that might involve different costs for the PES in function of the decision model. The disadvantage is that they can only be estimated for profiling models that return a value measured at ordinal level.

On classification discrimination, we want to estimate three metrics: accuracy, precision, and sensitivity (James et al., 2021). Denoting with TP the number of true positives, with TN the number of true negatives, with FP the number of false positives, and with FN the number of false negatives, they are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

Accuracy just looks at correct predictions, giving the same importance to events (LTU) and to non-events (not LTU). Sensitivity concentrates on the proportion of events that were correctly predicted, which seems closer to the interest of PES. Nevertheless, an obsession with sensitivity might provoke to wrongly classify too many people as positive, which is what precision reflects[79]. The prioritization of one metric over the other is intimately ligated to the decision model, so the PES should make the definitive decision on it. If the PES considers that the cost of not intervening in a jobseeker that will finally be LTU is larger than the cost of doing it in a not-LTU, it should give more importance to sensitivity. The opposite situation should pay more attention to the false alarm rate. For example, think in a very expensive intervention: giving the treatment to a person that will not be LTU might have a cost bigger than not providing it to a person that will be LTU, so precision turns crucial.

To compare the value of these metrics on classification discrimination, we need to select a threshold to transform scores into classes both for the rule-based and the statistical models. Rule-based profiling (Q-G and Q-S) departs from a non-

[78] Recall is a synonym of sensitivity. All the confusion matrix statistics will be defined in the next paragraph.

[79] Two additional metrics are also estimated and presented in the Appendix G: Kappa (a chance-corrected accuracy) and the false alarm rate ($FP/(FP + TN)$).

statistical selection of parameters, so the probability cutoff should not be an exception. Through this threshold we transform both multi-valued variables to dummies, with one value indicating the positive prediction of the event (or high risk of LTU) and the other the negative prediction of the event (or low risk of LTU). In the multi-class profiling (Q-G), the episodes assigned to the groups with the most intense treatment are codified as predicted events (SOC, 2016)[80]. In the score profiling (Q-S), we propose two scenarios. On one side, we classify as predicted event those episodes whose score is inside the top 50 % of possible values of the variable. Being this the standard practice with probability models, we denote this transformation as Y_{S50}. On the other side, we code as predicted event those episodes with a score in the top 25 % of possible values. This is a stricter rule, which would fit with a scenario of more restricted budget. Formally, these transformations are defined as follows[81]:

$$Y_{S25} = \begin{cases} Yes & S \leq 0.25\ 139 \\ No & S > 0.25\ 139 \end{cases}$$

$$Y_{S50} = \begin{cases} Yes & S \leq 0.5(139) \\ No & S > 0.5(139) \end{cases}$$

$$Y_{G} = \begin{cases} Yes & G \in \{\text{C1}, \text{C2}, \text{D}\} \\ No & G \notin \{\text{C1}, \text{C2}, \text{D}\} \end{cases}$$

For the statistical profiling, we use four models that return an estimated probability of LTU (Y). We need to convert this multi-valued variable to a binary one, so we present two possible transformations that correspond to different priorities that might have a PES (Loxha & Morgandi, 2014).

Classification policy A prioritizes to maximize the number of events detected. This rationale aligns with an interpretation of probabilities as propensities, thinking on binary events as the consequence of a latent variable model (Fox, 2015). This latent variable would have a cutoff that once is crossed, would normally cause the event to occur. Thus, the probability threshold (C) would be a parameter that can be estimated. Like a tuning parameter, we approximate it through cross-validation in the evaluation subset, so it is denoted as $C^{(eva)}$. This procedure tries to maximize Youden's $J = \frac{TP}{TP+FN} + \frac{TN}{TN+FP} - 1$, a function that

[80] They are the unique groups that receive an individual intervention as first action. The list of first and second interventions is available in Appendix A.

[81] S and G are defined in Appendix A.

balances the rate of accurate predictions of events (sensitivity) and the rate of accurate predictions of non-events (specificity).

Classification policy B prioritizes the compliance with a restricted budget to fund interventions. The idea is that, if the decision model is fixed to treat everyone profiled as positive (high risk of LTU), it might be actually impossible to meet this expense. The rationale is to set in advance the number of positives, since this will imply the number of treated jobseekers. With this classification policy, we code as event only those jobseekers whose estimated score is located at least at the nineth decile of the predicted scores found in the evaluation subset ($\mathbb{D}_9^{(eva)}$). The underlying plan is to treat only those at the highest decile of risk, but in the daily service we cannot wait to gather all the sample of candidates to then estimate the nineth decile. With the function we have presented, we estimate the point at which the cutoff of this nineth decile might be[82]. In a sense, this classification policy conflates the profiling and the decision stages, so it may be understood as a provisional solution until the full specification of an updated decision model.

Thereby, the two classification policies for the statistical profiling models are defined as follows:

$$Y_A = \begin{cases} Yes & Y \geq C^{(eva)} \\ No & Y < C^{(eva)} \end{cases}; \quad Y_B = \begin{cases} Yes & Y \geq \mathbb{D}_9^{(eva)} \\ No & Y < \mathbb{D}_9^{(eva)} \end{cases}.$$

The target parameters on calibration are two, from a less to a stricter interpretation of this concept. Mean calibration is a metric that relates the proportion of observed events to the proportion of expected events, denoting it by O:E (Van Calster et al., 2019)[83]. Moderate calibration is a more fine-grained statistic and refers to the integrated calibration index of Austin & Steyerberg

[82] Bach et al. (2023) calculated the decile using the test subset. If we understand the test subset as the data we would gather during the actual deployment of a profiling model, this profiling occurs in a one-to-one service provision. This means that in real life we would need in advance an estimation of the cutoff without the need to gather the whole test subset. This is our reason to use a subset of the data collected before the moment of the test subset. Of course, this assumes that the distribution of scores in the years of the evaluation subset barely changes regards the distribution of scores in the test subset.

[83] To talk about calibration, we adopt the classical interpretation of probability as relative frequencies in the limit, so we speak on probabilities and proportions as synonyms.

(2019). In short, this index is a weighted mean of absolute differences between the diagonal line of perfect calibration and the calibration curve. This calibration curve is a smooth representation of the scatterplot of the expected probability for each unit in the sample through a restricted cubic spline of five knots. To estimate these calibration metrics on the Q-S score (S_{it}), we need to transform it to a probability measure. Moreover, we have to reverse it, since it originally collected the probability of employment instead of unemployment. Thereby, we finally use $S_{it} = \frac{139 - S_{it}}{139}$.

Once we have determined how we transform the multi-valued outputs of the different models to a binary response (event vs. not event), we are almost ready to compare the predictions. To have a fair comparison, we need to use the same observations for each model. For comparisons between statistical models, this is easily done using the same test subset. For comparisons between (deployed) rule-based and our statistical models, we use the restrictive test subset but calculating the values of the time-variant variables at the moment the Q profiling was executed.

3.2.3. Model selection and interpretation

The selection of one model instead of another for deployment is not a trivial decision once we know that different models may achieve almost the same performance (Breiman, 2001a). The problem is that models with equivalent performance may produce different unit predictions (classifications), something that has been called model discrepancy (Marx et al., 2020) or model multiplicity (Black et al., 2022). Thereby, the greater the discrepancies between models, the larger the repercussions of adopting one model instead of the other. Following Geirhos et al. (2020), we employ Cohen's Kappa to estimate how prevalent is model discrepancy in our context. We had previously used Cohen's Kappa to compare a true value with a predicted value from one model, but we can also use it to estimate the chance-corrected coincidence between two models. Notice that if we want to compare model m with model b, $Kappa = (v_{obs_{m,b}} - v_{exp_{m,b}})/(1 - v_{exp_{m,b}})$. On one side, $v_{obs_{m,b}} = e_{m,b}/n$ represents the classical accuracy by relating the number of equal responses ($e_{m,b} = \sum_{it} \mathbf{1}(Y_m = Y_b)$) to the total number of units (n). On the other side, $v_{exp_{m,b}} = \Pr\big[Y = Yes\big]_m \Pr\big[Y = Yes\big]_b + \Pr\big[Y =$

$No\big]_m \Pr\big[Y = No\big]_b$ represents a binomial process in which the classification of each model is independent of the classification of the another.

To select which model to deploy, we may also consider how well reacts each of them to different stress tests (D'Amour et al., 2022). Stress tests evaluate the predictive performance of a model by introducing certain changes on the inputs of the functions related to desired properties of a model. Firstly, we have a technique called shifted performance evaluation that reviews the predictive performance of each model when introducing a sample with a different multivariate distribution of predictors to that of the training subset. This is what we do when we evaluate the performance of the models in the restrictive test subset, since this is a sample of those profiled with the deployed model. Secondly, stratified performance evaluations study if the predictive performance is equivalent in different strata of our population. It is a reduced version of some of the algorithmic fairness metrics collected in Mitchell et al. (2021). Following the documents of the Catalan PES (SOC, 2023), we focus on two strata: older jobseekers (over 45 years old) and older female jobseekers.

Another dimension to consider when choosing a model is its interpretability, although the understanding of what determines in general the predictions is per se an interesting question. Social scientists usually use the odds ratio or the average marginal effect (also known as average risk difference) to approximate the importance of each variable in a generalized linear model with a non-identity link function. In the machine learning literature (Molnar, 2022), researchers differentiate interpretation statistics that may be applied only to some models (model-specific) from others that are suitable for many types of modelling techniques (model-agnostic). Odds ratios are a model-specific metric, and average marginal effects use as building-block the average of unit predictions, which is equivalent to the model-agnostic method called partial dependence plot in the machine learning literature (Arel-Bundock et al., forthcoming).

Model-agnostic metrics are preferable if we want to compare results from different modelling techniques, so odds ratios are not an option. We have opted for another model-agnostic metric, the permutation-based variable importance (Fisher et al., 2019), for two reasons. On the one hand, this statistic directly measures the change in the ROC-AUC caused by the inclusion of each variable. Whereas average marginal effects focus on how the outcome varies when changing the value

of one predictor, we are actually interested in how the predictive performance is altered. Thus, the average marginal effect only looks at the predicted value, whilst the permutation-based metric also attends the true value. On the other hand, this is the statistic that has been used in previous research like Bach et al. (2023), which guarantees an easy comparison between findings. Specifically, we carried out ten rounds of permutation using a random sample of 10,000 observations to avoid an excessive computational burden.

4. Results

4.1. Performance comparison

In this section, we show the findings on the performance of the different models to predict LTU. Firstly, we compare the discrimination and calibration of our statistical models. Secondly, we use the restrictive test subset to contrast the performance of the current rule-based models versus the statistical models we have developed.

We start looking at the metrics that do not require to fix a unique probability threshold, that is, ranking discrimination and calibration. Table 4 reports the area under the ROC and the PR curves joint with the calibration statistics for the unpenalized and penalized logistic regression (LR and PLR), the random forest (RF), and the gradient boosting model (GB). Consistent with previous findings for Germany (Bach et al., 2023), tree-based models rank better than linear models with respect to ROC-AUC and PR-AUC, but the upgrades are not enormous. For both statistics, the gradient boosting model achieves the highest value. The next best model is the random forest, which makes sense thanks to the great flexibility of machine learning techniques. In any case, it is good news that the four statistical models are far from a ROC-AUC of 0.5, which is the expected value that we would have just by chance. In relation with the international literature, studies like Desiere & Struyven (2021) for Belgium and van den Berg et al. (2024) for Germany attained a slightly lower ROC-AUC defining LTU as six months of unemployment. Bach et al. (2023) used our definition of LTU for Germany and got very similar results.

Table 4 - Ranking performance of final models in the test subset (2022).

	ROC-AUC	*PR-AUC*	*O:E*	*ICI*
LR	0.742	0.398	0.497	0.170
PLR	0.745	0.396	0.503	0.166
RF	0.758	0.419	0.531	0.147
GB	0.763	0.433	0.603	0.110

Concerning calibration, Table 4 presents the estimates of the O:E and ICI statistics for each model. Remind that the O:E relates the observed average probability of LTU to the expected average probability of LTU as predicted by each model. This means that for this metric the closer to 1, the better. The gradient boosting model is the one that gets the most reliable predictions, since it is the one who is nearest to 1. Notice that all the models overestimated the average LTU probability, which might be a signal of overfitting. The ICI is a stronger version of calibration, showing the distance between the observed and the predicted values at the whole range of the function. As it is calculated as a residual, it is good news to be close to 0. We see that the gradient boosting model is again the function that attains the best performance at this metric. Specifically, the average residual of this model is of 11 percentual points. As far as we know, our research is the first in the jobseeker profiling literature to incorporate this detailed calibration metric.

The next question is whether our statistical models do better than the current rule-based procedures. After taking the restrictive test subset, Table 5 shows the ranking performance of the rule-based Q-S model joint with the results for our statistical profiling. Regarding discrimination, all statistical models attain higher values of ROC-AUC and PR-AUC than the rule-based procedure. The random forest is now the one who gets the best estimates, showing robustness to its evaluation with a sample not coming from the same distribution of the training subset. The Q-S model achieves a modest value both for the ROC-AUC and the PR-AUC. In this case, employing the Q-S function, if we randomly took one episode from a bag of actual events and another episode from a bag of actual non-events, the probability that the actual event had a larger score is 59.3 %[84]. Such discrimination ability is remarkably lower than the 73.5 % of ROC-AUC found

[84] This is the interpretation of the ROC-AUC as the probability of concordance (the c statistic).

for the random forest. This difference in performance is even bigger if we compare the area under the precision-recall function for the Q-S and the random forest models.

Table 5 - Ranking performance of final models in the restrictive test subset.

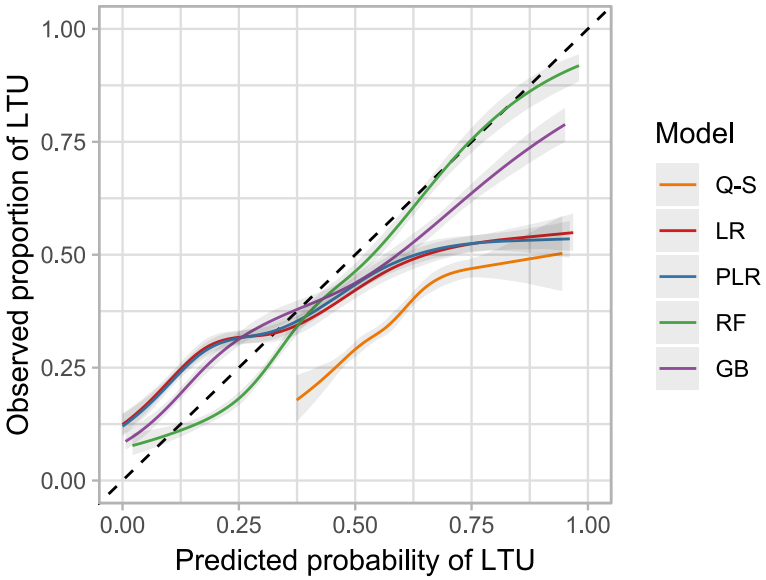
	ROC-AUC	*PR-AUC*	*O:E*	*ICI*
Q-S	0.593	0.430	0.609	0.223
LR	0.646	0.480	0.937	0.123
PLR	0.648	0.479	0.943	0.118
RF	0.735	0.603	0.908	0.037
GB	0.696	0.557	1.015	0.067

The advances are more remarkable when we compare the calibration of rule-based and machine learning models. The last column of Table 5 collects the ICI estimated for each model, showing that the random forest presents a tiny average residual, i.e. it deviates in average only 3.7 p.p. Notice that this is a sixth part of the ICI that presents the Q-S model, which appears poorly calibrated according to this index. To have a better idea on the calibration across the whole range of the function, Figure 1 plots each calibration curve. If we look at the mean calibration indicator (O:E), the gradient boosting model notably improves the value of the rule-based model by achieving almost exact coincidence (1.015). Felgueroso et al. (2018) did estimate the mean calibration for their statistical profiling model with Spanish data and attained almost the same result that we got with our best model (O:E = 0.999).

Despite both machine learning models outperform the rule-based procedure, it is noteworthy that the gradient boosting model had in general worse ranking metrics than the random forest in the restrictive test subset. A possible reason is that this dataset is too different from the training or the test subsets, i.e. that their distributions are not similar. To check this hypothesis, we compared the first moments of the predictors in the restrictive test and test subsets. We detected that the largest differences were in the proportions of jobseekers whose last employment was not temporary (11.4 p.p. more in the test), involved commuting

(9 p.p. extra), or required tertiary education (7.2 p.p. more)[85]. Thereby, we see that this shifted performance evaluation effectively used as inputs some variables with different distributions. In this case, the gradient boosting model is less robust to changes in the distributions of the covariates. Considering this evidence, according to ranking discrimination and calibration, the random forest seems the "best" statistical profiling model.

Figure 1 – Calibration curves in the restrictive test subset for each model.



We continue examining the metrics that need to establish a classification threshold to code a case as a predicted event, the so-called classification performance statistics. Table 6 reports our estimates of these metrics using the test subset. According to policy A, the classification that employs an optimized cutoff, there are small differences in terms of the three statistics for the LR, the PLR, and the RF. On accuracy and precision, there are small upgrades with the random forest respect to the logit models, finding the largest advances when using a gradient boosting model. On sensitivity, a reasonable target of PES, the random forest excels by attaining a value of 0.793. Note that compared with the unpenalized logit model, the random forest achieves increases both in sensitivity and in

[85] Complete results on this comparison are available in Appendix G. The comparisons do not include the dummies of missingness.

precision, which is positive regarding the efficiency of services. In fact, our random forest gets better results in sensitivity than the model developed by Felgueroso et al. (2018) for Spain, who has a sensitivity equal to 0.682[86]. On the contrary, the gradient boosting machine shows a decrease in sensitivity compared with the random forest, which could be a signal of overfitting.

Table 6 – Classification performance in the test subset based on different policies.

	Accuracy	Precision	Sensitivity
Policy A			
LR	0.572	0.251	0.794
PLR	0.580	0.254	0.790
RF	0.595	0.263	0.793
GB	0.667	0.296	0.729
Policy B			
LR	0.782	0.377	0.483
PLR	0.782	0.378	0.482
RF	0.830	0.481	0.323
GB	0.826	0.469	0.362

According to policy B, the classification that considers budget limitations, the values of accuracy are in all cases larger than those of the previous policy. Due to the remarkable reduction of the sensitivity of all models, we can safely conclude that this accuracy increase comes from detecting a greater proportion of non-events. This is a consequence of the initial restriction that policy B establishes on the number of jobseekers that may be coded as predicted events. In cases of restricted budgets, an alternative to these two policies could be to avoid the use of classification thresholds based on binary variables and split the variable of LTU probability in an ordered polytomous variable. Establishing five categories of risk

[86] Felgueroso et al. (2018) used a slightly different process of model validation, since they selected the classification threshold with the same data that they estimated sensitivity (the test subset). In our case, we fix it in advance with the evaluation subset. This precludes a too optimistic performance evaluation. The reasons are that: (1) if the test subset took a different year of implementation, it might have a different multivariate distribution than the evaluation subset, and (2) in real implementations, we need to establish the classification threshold in advance to allocate a jobseeker, instead of waiting until we have collected all the data for the year.

instead of two might help the PES to attend each segment of jobseekers only if the present budget allows to treat the next level of risk.

The next step is to compare the classification performance metrics of these models with those from the rule-based procedures currently in use. They are shown in Table 7, including two classification rules for Q-S (the function that outputs a score) and one for Q-G (the function that returns a class). In the first panel, we appreciate that Q-S achieves a remarkable sensitivity when placing its cutoff at the 50 % (Q-S50). If we dig a little deeper, we will see that the problem is that this high sensitivity is attained through a very unprecise classification, as confirmed by the substantial false alarm rate presented in Appendix G. When locating the threshold at the top 25 % (Q-S25), the precision increases and the sensitivity almost disappears. If we look at the rule-based model that outputs a class (Q-G), we see a better level of accuracy and a small value of sensitivity. This suggests that this model only performs relatively well to detect non-events.

The metrics of the statistical models generally outperform those of the rule-based procedures, and between these statistical devices the patterns found with the restrictive test are very similar to the ones encountered with the full test. In sum, statistical models that optimized its classification threshold present larger specificity (policy A), whereas those that prioritized budget limitations have higher accuracy (policy B). If we assume that detecting LTU is the main objective of PES, the random forest under policy A executes this task better than any other statistical or rule-based model. The additional advantage is that this statistical model is more precise with its classification. Therefore, it outperforms the discrimination ability of the rule-based model Q-S50. If the PES aims to achieve a more balanced approach between predicting events and non-events, the gradient boosting model could offer nearly the same accuracy as Q-S25 while increasing its sensitivity by more than sevenfold.

Table 7 – Classification performance in the restrictive test subset based on different policies.

	Accuracy	Precision	Sensitivity
Q-S25	0.641	0.482	0.090
Q-S50	0.454	0.381	0.852
Q-G	0.605	0.395	0.204
Policy A			
LR	0.562	0.432	0.729
PLR	0.567	0.436	0.735
RF	0.579	0.452	0.860
GB	0.613	0.472	0.730
Policy B			
LR	0.653	0.517	0.404
PLR	0.653	0.516	0.406
RF	0.699	0.619	0.402
GB	0.679	0.578	0.367

Note: N of the restrictive test subset = 11,082.

The joint analysis of discrimination and calibration illustrates that the random forest model under policy A improves the rule-based Q-S50 in all performance statistics. We have paid special attention to this last rule-based model because it is the one that excels in terms of sensitivity. However, any statistical model (under policy A or B) does better than the qualitative rule-based profiling (Q-G) in the classification metrics focused on the events. The random forest is particularly outstanding due to its extraordinary calibration, while Q-S suffers from the unreliability of its numerical predictions. This statistical model also achieves a remarkable sensitivity (0.860) and upgrades regarding precision and accuracy, which might promote a more efficient allocation of treatments.

4.2. Model similarity and interpretation

The previous results show that when we use the test subset the performance metrics of the different statistical models are not extremely different in many cases. This section measures the phenomenon of model similarity in a more formal

way, looking both at the prevalence and the consequences of these resemblances. Lastly, we also investigate if the most important predictors are the same for each modelling technique.

Table 8 – Kappa coefficients between predictions of different models in the restrictive test subset.

	Q-S25	Q-S50	Q-G	LR	PLR	RF	GB
Q-S25	1						
Q-S50	0.036	1					
Q-G	0.044	0.051	1				
Policy A							
LR	0.016	0.014	0.018	1			
PLR	0.015	0.017	0.017	0.925	1		
RF	0.018	0.031	0.025	0.576	0.587	1	
GB	0.018	0.011	0.028	0.703	0.715	0.671	1
Policy B							
LR	0.014	0.011	0.029	1			
PLR	0.007	0.009	0.029	0.949	1		
RF	0.031	0.025	0.037	0.587	0.586	1	
GB	0.017	0.019	0.046	0.723	0.725	0.732	1

Table 8 estimates the degree of coincidence among models through the Kappa coefficients. Considering that this statistic is located in the interval [0, 1], we see that the agreement between rule-based and statistical models is very small. Our explanation is that the rule-based functions mostly classify jobseekers randomly, that is, by chance[87]. This makes us conclude that the disagreement shown by these Kappa do not mostly come from rule-based models collecting a different data generating process. If we concentrate on the statistical models, we may extract two main ideas. First, LR and PLR models show an almost total agreement in their classifications, which makes sense if we see that PLR models incorporate

[87] We show evidence of this statement in Appendix G, where we present for each profiling model the Kappa coefficient that compares the predicted and the actual value. Q-S50 presents a chance-corrected accuracy equal to 0.067.

small penalization terms after cross-validation (see Appendix F). Second, the two tree-based models have an intermediate degree of agreement, particularly under policy A. This prompts us to examine the effect of selecting one model over the other.

Through the following stress-tests we can appreciate how robust is each model to a slightly different population and how much varies the performance when we concentrate on some target groups. Firstly, the shifted performance evaluation was implemented when we used the restrictive test subset to estimate discrimination and calibration metrics. Remind that the restrictive test subset includes a moderately different population because it is made up of those individuals that wished to participate in ALMP or were profiled as a requirement of a non-contributory benefit. Our conclusion for that analysis was that, if detecting LTU is the priority of the PES, the random forest had better discrimination ability and higher moderate calibration. Secondly, the stratified performance evaluation on older and older female jobseekers is available in Table 9. The results show that the random forest attains higher sensitivity than the gradient boosting model in both groups. Nonetheless, it is important to notice that linear models (LR and PLR) register almost the same sensitivity, although they do it with less precision and less accuracy. This should also be considered joint to their easy interpretation when deciding which model to implement. As we mentioned before, another decision dimension that must be clarified are the prediction priorities of the PES, which is usually linked to the costs of each treatment. If the prediction of non-events is more important due to the high cost of LTU treatments, the gradient boosting model shows the best accuracy and the lowest sensitivity, which indicates a high true negative rate.

Table 9 – Classification performance in two strata of the test subset.

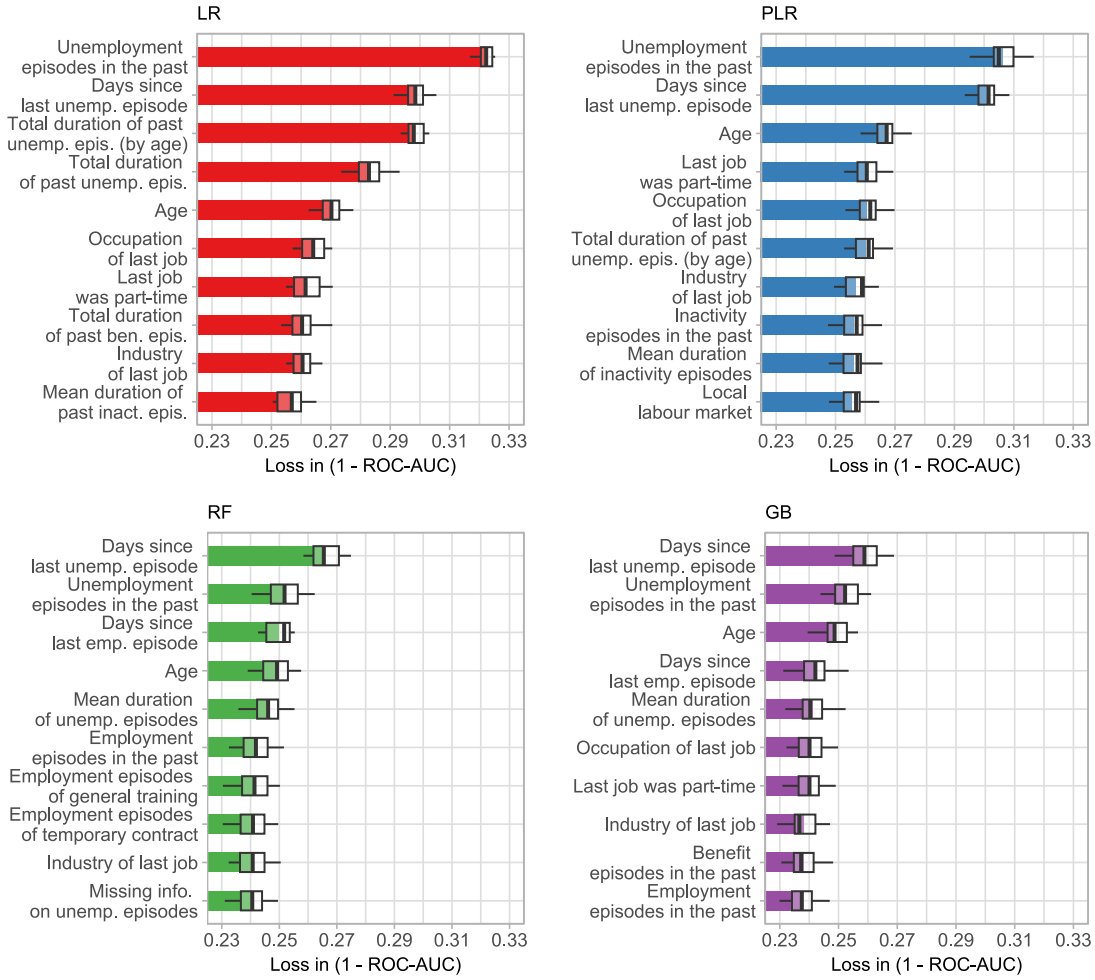
	Accuracy	Precision	Sensitivity
Older jobseekers			
LR	0.551	0.277	0.826
PLR	0.547	0.276	0.838
RF	0.587	0.295	0.825
GB	0.653	0.328	0.769
Female and older			
LR	0.543	0.279	0.847
PLR	0.537	0.278	0.859
RF	0.603	0.309	0.837
GB	0.665	0.344	0.794

Note: N of older jobseekers = 20,793. N of older female jobseekers = 11,452. For simplicity, we only include results for policy A.

To end up, the analysis of variable importance helps us to understand which are the predictors whose presence might explain the superior performance of statistical models[88]. For each statistical model, Figure 2 represents the ten most important predictors. Remind that this is operationalized through the loss in the ROC-AUC caused by permutating the values of such covariate. We see that the four models coincide when flagging the two most relevant predictors of LTU: the number of days since the last unemployment episode and the number of unemployment episodes experienced in the past. Age is another predictor that stands out in the four models. This is something consistent with the results of Felgueroso et al. (2018) for Spain. Tree-based models point up two additional remarkable predictors of LTU: the number of days since the last employment episode and the average duration of unemployment episodes in the past. This is not surprising, since previous studies like Bach et al. (2023) and McGuinness et al. (2022) also detected that labour market histories are key to predict LTU in Germany and Ireland.

[88] We have also used other technique to interpret the variable importance called global surrogate model. This method simply fits a decision tree to the unit predictions of the random forest trying to assess which covariates collect most part of the variance of the predictions vector. It is available in Appendix F.

Figure 2 – Top-10 variable importance of final models.



Note: The size of the bar shows the value of the permutation statistic. It is an average across permutation rounds, so the boxplot represents the variability among rounds.

5. Conclusion

In this chapter, we have compared the performance of rule-based and statistical models for jobseeker profiling. Using as case of study the PES of Catalonia (Spain), we have contrasted the predictions of its currently implemented rule-based model with the predictions of different statistical models of LTU. We have found that our statistical models remarkably improve the rule-based procedures with respect to discrimination (ROC-AUC: 0.735 vs. 0.593) and with respect to calibration (ICI: 0.037 vs. 0.223). Moreover, we have shown that machine learning

models attain greater performance values than classic regression models, particularly with respect to calibration. These models represent the first machine learning algorithms developed and validated to predict LTU with Spanish microdata. In comparison with the gradient boosting model, the random forest handles better the covariate shifts and gets higher sensitivity for two subpopulations targeted by the Catalan PES (older and older female jobseekers). Furthermore, the statistical models have indicated two covariates that are not currently used in the profiling function although they seem good predictors of LTU: the number of days since the last unemployment episode and the number of past unemployment episodes.

Our results are consistent with previous evidence of the jobseeker profiling literature, but also proposes new points of view. Following former studies, we also detect the relevance of labour market histories to predict LTU (Gabrikova et al., 2023; McGuinness et al., 2022). Nevertheless, prior research had concluded that machine learning models do not involve remarkable improvements in performance over classic techniques like logit models (Bach et al., 2022; Desiere et al., 2019). Our key result is that this statement only holds if we concentrate on the discrimination dimension of performance. Discrimination is just one side of the coin and if we focus on the other, i.e. calibration, our machine learning models outperformed logistic regression. Although it is not often attended in the jobseeker profiling literature, we have argued that calibration is relevant in PES applications because it informs the jobseeker on the uncertainty of a binary prediction.

The research has certain drawbacks that should be taken into account, in some cases because they open lines of future research. First, we do not have information on employment episodes as detailed as Bach et al. (2023), mainly because our datasets do not register the actual end dates of the contracts. As a following step, researchers might try to get social security data to use the intervals of affiliations. Second, our training data might not exactly represent the distribution of all jobseekers in Catalonia. The reason is that we have utilized a sample of the individuals registered in the Catalan PES, but such registration is not compulsory in all cases in the Spanish context. Jobseekers without enough labour market experience like young people are not forced to be in the PES system, so they might be underrepresented. Third, when we optimized the classification threshold of our statistical models, we did it under the assumption that the cost of a false positive equals the cost of a false negative. This profiling option should be

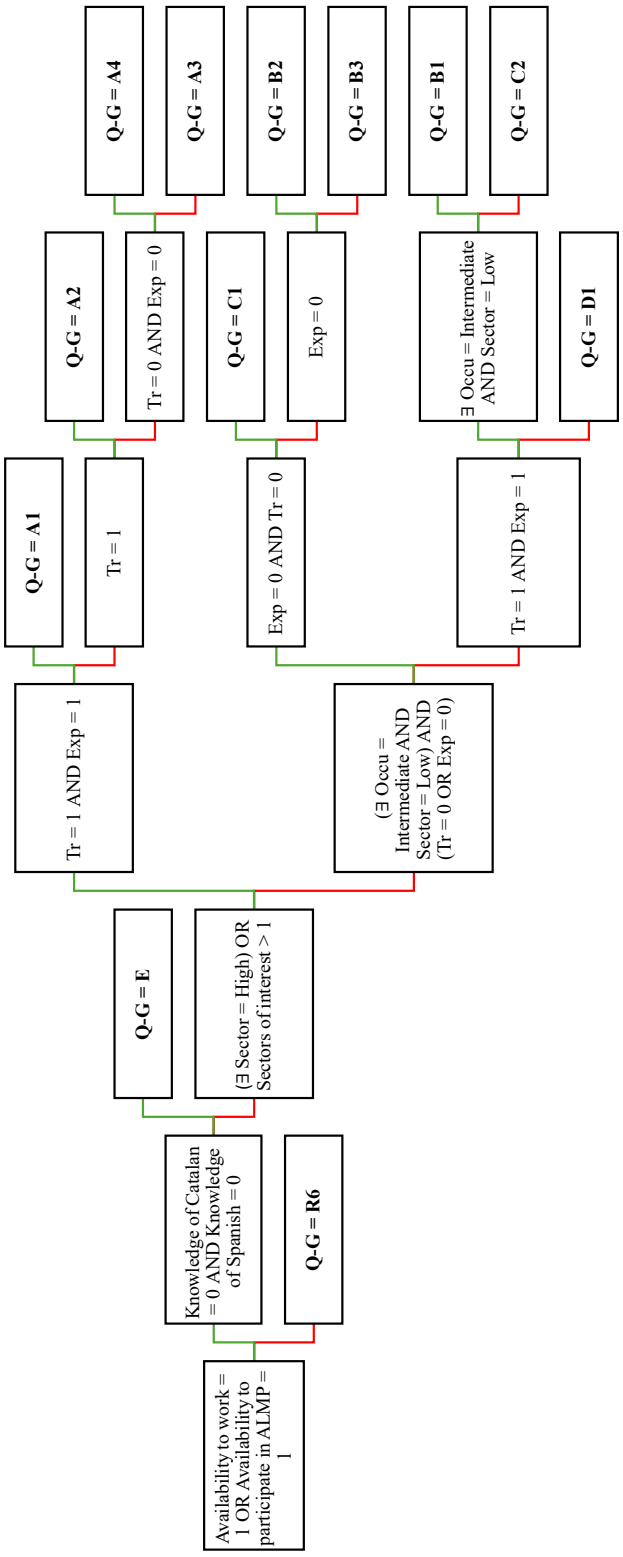
informed by the decision of the PES. Thereby, a promising line of future work is the development of a decision model for all the available programs. Notice that the separation between decision and prediction helps to address the problems of discretion discussed in Gallager & Griffin (2023) on the selection of the performance metric. Fourth, the next step in the model validation process should be an exhaustive fairness assessment of the different models. In this research, we have examined if the performance of the models was equivalent for two target subpopulations. Nonetheless, future research might evaluate if error rates are similar for other sensitive social groups.

Beyond all these shortcomings, this chapter has emphasized the benefits of implementing a statistical instead of a rule-based profiling model in PES. Moreover, we have also demonstrated the advantages of using machine learning models in terms of discrimination and also calibration for a better design of jobseeker allocation systems.

Appendix

Appendix A. Current profiling and decision models of SOC

Figure A1 – Rule-based profiling model Q-G



Note: Own elaboration based on SOC (2016). This decision tree was inferred from the documents provided, so it must be taken with caution. The green line indicates the path if the value is TRUE, the red line if the value is FALSE. It represents the values considered to assign an individual to a treatment, which are the so-called pre-collectives. *Training* indicates that training is enough (1) or not enough (0), denoted as $Tr := \{1,0\}$. *Experience* indicates if experience is enough (1) or not enough (0), denoted as $Exp := \{1,0\}$. The variable employability of the occupation of interest is represented as $Occu := \{High, Intermediate, Low, \emptyset\}$, and was originally denoted {"*Viable*", "*Moderado*", "*En retroceso*", "*No definida*"}. The variable employability of the sector of interest is represented as $Sector := \{High, Low\}$, and was originally denoted as {"*No en retroceso*", "*En retroceso*"}.

Table A2 – Rule-based profiling model Q-S

$\text{Q-S} = (S_{ito_1}, S_{ito_2}, S_{ito_3})$. $S_{it} = \text{mean}(S_{ito_1}, S_{ito_2}, S_{ito_3})$.

$$\begin{aligned} S_{ito} = {} & \lambda_1 occupation_{ito} + \lambda_2 sector_{ito} + \lambda_3 regexperience_{it} + \lambda_4 irrexperience_{it} \\ & + \lambda_5 regexperience_{ito} + \lambda_6 irrexperience_{ito} + \lambda_7 education_{it} \\ & + \lambda_8 comptrain_{it} + \lambda_9 driving_{it} + \lambda_{10} closed_{ito} + \lambda_{11} search_{it} \\ & + \lambda_{12} languages_{it} + \lambda_{13} digitalskills_{it} + \lambda_{14} willingness_{it} \\ & + \lambda_{15} communication_{it} + \lambda_{16} interpersonal_{it} \end{aligned}$$

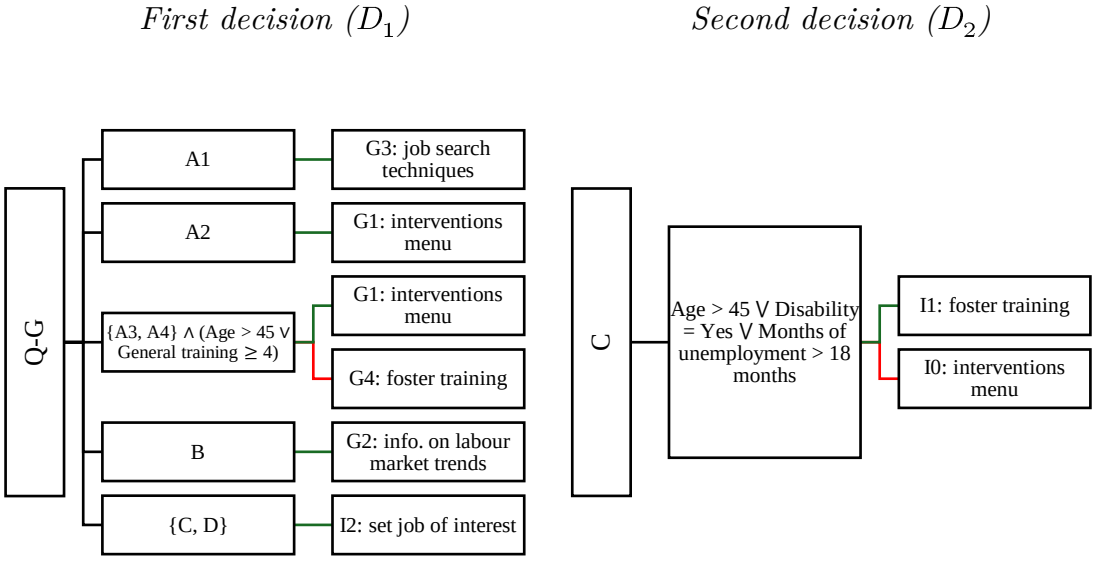
Variable (index L)	*Value*	λ_L
$occupation_{ito}$	$x =$ level 1	$\lambda_1 = 0$
$(L = 1)$	$x =$ level 2	$\lambda_1 = 5$
	$x =$ level 3	$\lambda_1 = 6$
	$x =$ level 4	$\lambda_1 = 7$
	$x =$ level 5	$\lambda_1 = 10$
$sector_{ito}$	$x =$ emergent	$\lambda_2 = 10$
$(L = 2)$	$x \in$ {more than one, normal}	$\lambda_2 = 5$
	$x \in$ {declining, missing value}	$\lambda_2 = 0$
$regexperience_{it}$	$x >$ 60 months	$\lambda_3 = 10$
$(L = 3)$	36 months $< x \leq$ 60 months	$\lambda_3 = 7$
	18 months $< x \leq$ 36 months	$\lambda_3 = 5$
	6 months $< x \leq$ 18 months	$\lambda_3 = 3$
	0 months $\leq x \leq$ 6 months	$\lambda_3 = 0$
$irrexperience_{it}$	$x >$ 60 months	$\lambda_4 = 2$
$(L = 4)$	6 months $< x \leq$ 60 months	$\lambda_4 = 1$
	0 months $\leq x \leq$ 6 months	$\lambda_4 = 0$
$regexperience_{ito}$	$x >$ 60 months	$\lambda_5 = 10$
$(L = 5)$	36 months $< x \leq$ 60 months	$\lambda_5 = 7$
	18 months $< x \leq$ 36 months	$\lambda_5 = 5$
	6 months $< x \leq$ 18 months	$\lambda_5 = 3$
	0 months $\leq x \leq$ 6 months	$\lambda_5 = 0$
$irrexperience_{ito}$	$x >$ 60 months	$\lambda_6 = 2$
$(L = 6)$	6 months $< x \leq$ 60 months	$\lambda_6 = 1$
	0 months $\leq x \leq$ 6 months	$\lambda_6 = 0$

Variable (index L)	*Value*	λ_L
$education_{it}$	$x \in \{0,$ 1nr, missing value, 25Bnr$\}$	$\lambda_7 = 0$
$(L = 7)$	$x \in \{1,$ 25B, 24nr, 35Bnr$\}$	$\lambda_7 = 2$
	$x \in \{24, 35B, 44nr, 45nr, 55Bnr, 54nr\}$	$\lambda_7 = 4$
	$x = \{44,$ 45, 54, 55B, 55nr, 75Bnr$\}$	$\lambda_7 = 6$
	$x \in \{6nr,$ 7nr, 8nr, 55, 75B$\}$	$\lambda_7 = 8$
	$x \in \{6,$ 74, 75, 8$\}$	$\lambda_7 = 10$
$comptrain_{it}$	$x \geq 80$	$\lambda_8 = 10$
$(L = 8)$	$x < 80$	$\lambda_8 = 0$
$driving_{it}$	$x =$ B1	$\lambda_9 = 10$
$(L = 9)$	$x \neq$ B1	$\lambda_9 = 0$
$closed_{ito}$	$x =$ (Yes $\wedge$ (Has it $\vee$ Studying it))	$\lambda_{10} = 5$
$(L = 10)$	$x \in \{$Recommended $\wedge$ Has it $\vee$ Studying it , No $\wedge$ (Has ISCED 2 $\vee$ Studying ISCED 2)$\}$	$\lambda_{10} = 4$
	$x \in \{$Recommended $\wedge$ Doesn't have it, No $\wedge$ Has not ISCED 2$\}$	$\lambda_{10} = 2$
	$x \in \{$Yes $\wedge$ (Doesn't have it $\vee$ Unknown), (Recommended $\vee$ No) $\wedge$ Unknown$\}$	$\lambda_{10} = 0$
$search_{it}$	$x = 0$	$\lambda_{11} = 10$
$(L = 11)$	$x = 1$	$\lambda_{11} = 5$
	$x = 2$	$\lambda_{11} = 3$
	$x = 3$	$\lambda_{11} = 0$
$languages_{it}$	$Cat \in \{$High, Middle$\} \wedge Sp \in \{$High, Middle$\}$	$\lambda_{12} = 10$
$(L = 12)$	$Cat \in \{$High, Middle$\} \wedge Sp \in \{$Basic, Null$\}$	$\lambda_{12} = 5$
	$Cat \in \{$Basic, Null$\} \wedge Sp \in \{$High, Middle$\}$	$\lambda_{12} = 4$
	$Cat \in \{$Basic, Null$\} \wedge Sp \in \{$Basic, Null$\}$	$\lambda_{12} = 0$
$digitalskills_{it}$	$x = 0$	$\lambda_{13} = 10$
$(L = 13)$	$x = 1$	$\lambda_{13} = 5$
	$x \in \{2, 3\}$	$\lambda_{13} = 0$
$willingness_{it}$	$x = 0$	$\lambda_{14} = 10$
$(L = 14)$	$x \geq 1$	$\lambda_{14} = 0$
$communication_{it}$	$x =$ Yes	$\lambda_{15} = 10$
$(L = 15)$	$x =$ No	$\lambda_{15} = 0$
$interpersonal_{it}$	$x =$ Yes	$\lambda_{16} = 10$
$(L = 16)$	$x =$ No	$\lambda_{16} = 0$

$education_{it}$ in ISCED-11 coding of educational attainment, with "nr" indicating that the credential has not been recognized in Spain and "B" indicating that the credential was obtained in the framework of labour market policies. "Cat" denotes

the Catalan language and "Sp" the Spanish language. ∨ is the Boolean operator for OR, ∧ is the Boolean operator for AND. Appendix D includes a description of each variable.

Figure A3 – Decision model for decisions one and two



Source: SOC (2016). The green line indicates the path if the value is TRUE, the red line if the value is FALSE. The tree starts with the value of Q-G or C (the so-called "pre-collective") and then links the decision, i.e. the intervention assigned. It also shows if such intervention is individual (I) or group-based (G) and the main objective of the action. The second decision is only determined by the rule-based profiling if individual was not assigned in Q-G to collectives C, D, or Z. In those cases, the second decision is not regulated by the model.

The interventions to assign in these decisions may be classified according to three variables (SOC, 2016):

1. *Number of participants*: one (individual) or more than one (group).
2. *Main objective*: training on job search techniques, foster adult training, information provision on labour market trends, set job of interest, or present the set of available interventions (the so-called "interventions menu").
3. *Place of implementation*: face-to-face or remote.

The intensity of the intervention may be defined according to different criteria. In this article, we have chosen to define individual actions as more intense than group actions, so $d \in \{\text{I0}, \text{I1}, \text{I2}\} > d' \in \{\text{G1}, \text{G2}, \text{G3}, \text{G4}\}$.

We have maintained the original abbreviations of the group interventions, but we have changed the abbreviations of the individual interventions to avoid confusions. I0 is known as "*Assessorament Polítiques Actives d'Ocupació*" (originally abbreviated as APAO), I1 is known as "*Assessorament Ocupacional*" (originally abbreviated as AO), and I2 is known as "*Orientació*" (originally abbreviated as O).

Appendix B. Review on performance of jobseeker profiling models

Country	*Model*	*Outcome*	*ROC*	*Sensit.*	*Prec.*	*Acc.*	*O:E*	*Source*
AT	ST	Labour market integration probability				[0.80, 0.85]		Desiere et al. (2019)
BE	ST, CW	MTU	0.76			0.67		Desiere et al. (2019)
BE*	ST	MTU	0.702			0.702		Desiere & Struyven (2021)
DK	ST	MTU				>0.60		Desiere et al. (2019)
FI*	ST	LTU	0.80					Viljanen & Pahikkala (2020)
DE*	ST	MTU	0.7			~0.63		Kunaschk & Lang (2022)
DE*	ST	LTU	0.777	0.29	0.372	0.846		Bach et al. (2023)
DE*	ST	MTU	[0.726, 0.735]	0.8		0.647	1.237	Van den Berg et al. (2024)
IE	ST	LTE				[0.70, 0.86]		Desiere et al. (2019)
IE	ST	LTU		0.752		0.777		McGuiness et al. (2022)
NL	ST	LTU				0.7		Desiere et al. (2019)
NZ	ST, RB	Lifetime income support costs, lifetime income support and staff costs	[0.63, 0.83]					Desiere et al. (2019)
SK*	ST	Categories of unemployment duration		0.7886	0.915	0.918		Gabrikova et al. (2023)
ES*	ST	LTE				0.682	0.999	Felgueroso et al. (2018)
GB*	ST	LTU	0.795	0.319	0.333	0.889		Matty (2013)

Note: Sensit.: sensitivity, Prec.: Precision, Acc.: Accuracy. CW: Caseworker-based, RB: Rule-based, ST: Statistical. MTU: medium-term unemployment ($\geq$ 6 months), LTU: long-term unemployment ($\geq$ 12 months), LTE: long-term employment (exit to employment within 12 months). ROC refers to the area under the ROC. This table only includes those models that were publicly validated with at least one statistic of discrimination or calibration. Classification metrics are only included if the author recommended or used at least one classification threshold. Note the differences between exit to employment *within* 12 months (at least once in the time interval) and exit to employment *after* 12 months (at the measurement time of month 12). For the results of Gabrikova et al. (2013), although their model uses four categories, here we present the metrics for the category "more than 12 months". (*) Rows with the asterisk indicate that, according to the source, the model has not been yet deployed in public employment services.

Appendix C. Classification of intermediation claims

Table C1 – Correspondence between causes of intermediation claims and type of episode

Code	*Description of the cause*	*Type*
1	Removal due to placement communicated with prior offer	E
2	Removal due to registration in the general Social Security system	E
3	Removal due to placement in the special self-employed regime	E
4	Removal due to placement communicated without prior offer	E
17	Removal due to the end of a collective dismissal file	U
19	Removal due to call of a seasonal permanent worker	E
25	Removal due to incomplete application	U
30	Suspension without intermediation due to temporary incapacity	I
31	Suspension without intermediation due to maternity/paternity, adoption, or foster care	I
32	Suspension without intermediation due to pregnancy with risk	I
35	Removal due to end of availability	I
36	Removal due to total permanent disability	I
37	Removal due to absolute permanent disability (major disability)	I
38	Removal due to retirement	I
39	Removal due to reaching the minimum retirement age	I
61	Removal due to other causes	I
62	Provisional removal due to untraceable applicant	U
70	Removal due to failure to appear before the managing entity	U
71	Removal due to failure to renew the application	U
73	Removal due to rejecting a suitable job offer	U
75	Removal due to refusal to participate in ALMP	U
100	Voluntary removal	U
102	Removal due to benefit exportation	I
103	Removal due to death	I
104	Suspension due to military service or alternative civilian service	I
105	Removal due to equalization	U
106	Suspension without intermediation due to preventive detention	I
107	Removal due to job placement declaration	E
108	Suspension without intermediation due to deprivation of liberty for fulfilling a sentence of applicants receiving benefits	I

109	Removal due to deprivation of liberty for fulfilling a sentence	I
110	Removal due to non-communication of the renewal of administrative authorization	I
114	Suspension without intermediation due to family obligations	I
120	Suspension without intermediation due to leaving the country	I
121	Suspension without intermediation due to attending training courses	U
122	Suspension with limited intermediation due to collective dismissal file or short-time working arrangements of suspension or reduction of working hours	U
125	Suspension due to cause 125	I
509	Removal due to accumulated benefit payment caused by return to the country of origin	I
530	Suspension due to temporary inability with intermediation	I
531	Suspension due to maternity/paternity, adoption, or foster care with intermediation	I
614	Suspension due to family obligations with intermediation	U
620	Suspension with intermediation due to leaving the country	I
621	Suspension with intermediation due to attending training courses	U
625	Suspension due to assignment to social collaboration work* with intermediation	E
626	Suspension due to deferred coverage with intermediation	E
627	Suspension due to deferred call with intermediation	E
700	Registration due to enrolment	U
701	Registration due to coverage of a vacancy (to be phased out)	E
702	Registration due to collective dismissal file	U
703	Registration due to correction of an erroneous removal	U
704	Registration with recovery of a period in a removal situation	U
706	Registration due to initial enrolment	U
707	Registration due to reactivation of suspension	U
708	Registration due to enrolment as employment intermediation	U
709	Registration as a jobseeker for other ALMP	U
710	Registration for ALMP prior to employment	U
711	Registration due to benefit resumption-compatibility	U

Note: E: part of an employment episode, U: part of an unemployment episode, I: part of an inactivity episode. (*) Social collaboration works are work-related activities, which includes both working and training, that may be required in exchange for the receipt of a benefit (Cremades Chueca, 2014).

Appendix D. List of predictors of Q, C and K

Table D1 – Predictors used in Q models.

Group	*Predictor*	*Label*
Job	Employability of the occupation of interest	$occupation_{ito}$
	Employability of the sector of interest	$sector_{ito}$
General employment experience	Months of experience in regular employment	$regexperience_{it}$
	Months of experience in irregular employment	$irrexperience_{it}$
Employment experience in the occupation	Months of experience in regular employment in the occupation of interest	$regexperience_{ito}$
	Months of experience in irregular employment in the occupation of interest	$irrexperience_{ito}$
General training	Level of education	$education_{it}$
	Number of hours of complementary training	$comptrain_{it}$
	Driving license	$driving_{it}$
Occupational training	Is the occupation closed (regulated) and has the jobseeker the credential?	$closed_{ito}$
Job search	Number of deficiencies in the knowledge and use of job search techniques	$search_{it}$
Language skills	Knowledge of Catalan or Spanish	$languages_{it}$
Digital skills	Number of deficiencies in ICT abilities	$digitalskills_{it}$
Transversal skills	Deficiencies in the willingness to learn	$willingness_{it}$
	Proper communication	$communication_{it}$
	Proper interpersonal relation	$interpersonal_{it}$

Source: Own elaboration based on screenshots of the Q software and SOC (2016).

Table D2 – Predictors used in the C function.

Group	*Predictor*
Used in formal allocation	Age
	He/she has a disability
	Duration of the unemployment episode
Not used in formal allocation	Sex
	He/she receives a benefit
	Geographical mobility
	Availability to work
	Availability to participate in ALMP
	Economic dependence

Source: Own elaboration based on SOC (2016).

Table D3 – Predictors used in our statistical models (K).

Group	*Predictor*
PLMP	Started unemployment during a benefit interval
	Number of benefit episodes (completed) in the past
	Total duration of previous benefit episodes
	Total duration of previous benefit episodes, scaled by age
	Mean duration of previous benefit episodes
ALMP	Total duration of employment subsidy participations
	Number of JSA/JSM participations in the past
	Number of training participations in the past
	Total durations of JSA/JSM participations in the past
	Total durations of JSA/JSM participations in the past, by age
	Total durations of training participations in the past
	Total durations of training participations in the past, by age
	Mean duration of training participations in the past (in days)
	Mean duration of JSAM participations in the past (in days)
Unemployment	Number of unemployment episodes in the past (inside the window)
	Total duration of unemployment episodes in the past (until the present episode, not included)
	Total duration of unemployment episodes in the past (until the present episode, not included), by age
	Mean duration of unemployment episodes until the present (until the present episode, not included)
	Days since last unemployment episode
Inactivity	Total duration of inactivity episodes
	Mean duration of inactivity episodes until the present
	Number of inactivity episodes in the past
Employment	Days since first employment (in the window)
	Days since (the beginning of) the last employment episode
	Days since (the beginning of) the last full-time employment episode
	Occupation of last job by major groups (63 categories)
	Last job was part-time
	Skill level required for last job (11 categories)
	Last job was temporary
	Industry of last job (22 categories)
	Commuted for last job
	Proportion of jobs with commuting in the past

	Number of employment episodes without any vocational training held in the past
	Number of occupations held in the past
	Number of employment episodes in the past
	Number of open-ended contracts in the past
	Number of temporary contracts in the past
	Maximum skill level required for past employment episodes (11 categories)
	Maximum skill level required for past employment episodes (5 categories)
	Minimum skill level required for past employment episodes (11 categories)
Socio-demographics	Sex
	Age in years when the episode started
	Maximum level of education
	Has a disability
	Local labour market (28 categories)
	National group (7 categories)
	Has a credential with field of education = xy (33 binary variables)
Missing blocks	Indicator of missingness on employment episodes in the past
	Indicator of missingness on unemployment episodes in the past
	Indicator of missingness on local labour market

Note: Qualitative variables that do not indicate the number of categories are binaries, so there are until three possible categories (yes, no, or missing). For the models that use regularization, this list is actually a list of candidate predictors.

Appendix E. Summary statistics

Table E1 – Summary statistics on sociodemographic qualitative variables.

	N	*%*
Sex		
Woman	153,424	52.412
Man	139,301	47.588
Maximum level of education		
0 Less than primary	4,965	1.696
1 Primary	8,137	2.780
24 Lower secondary – General	147,550	50.406
25 Lower secondary – Vocational	41	0.014
34 Upper secondary – General	27,382	9.354
35 Upper secondary – Vocational	40,327	13.776
55 Short-cycle tertiary – Vocational	32,071	10.956
66 Bachelor's	13,153	4.493
76 Master's	18,538	6.333
86 Doctoral	561	0.192
Disability		
No	273,733	93.512
Yes	18,992	6.488
National group		
Asia	788	0.269
EU, Northern America, and Oceania	281,292	96.094
Europe not EU	451	0.154
Latin America and the Caribbean	2,088	0.713
Northern Africa	6,352	2.170
Sub-Saharan Africa	1,749	0.597
Missing	5	0.002

Note: The categories related to the local labour market, the field of study and the level of study are not shown to simplify the exposition. The tables are available upon request.

Table E2 – Summary statistics on sociodemographic quantitative variables.

	Mean	*Min*	*Q1*	*Median*	*Q3*	*Max*
Age	46.183	16	40	46	52	64

Table F1 – Tuning grids.

Model	*Parameter*	*Candidate values*
Penalized logistic regression (PLR)	Amount of regularization (penalty)	**0.001**, 0.01, 0.1, 1, 10, 100, 1000
	Proportion of Lasso penalty (mixture)	0, **1**
Random forest (RF)	Number of predictors (mtry)	sqrt(# predictors), **log2(# predictors)**
	Minimal node size (min_n)	**1**, 5, 10
	Number of trees (trees)	500, **750**
Gradient boosting machine (GB)	Tree depth (tree_depth)	3, 5, **7**
	Number of predictors (mtry)	**sqrt(# predictors)**, log2(# predictors)
	Number of trees (trees)	250, **500**, 750
	Learning rate (learn_rate)	0.01, **0.025**, 0.05
	Proportion of sampled observations (sample_size)	**0.6**, 0.8

Note: In the parameter column, it is shown in parenthesis the name given in the R library {parsnip} to that parameter. The selected value is written in bold in the third column. The unpenalized logistic regression has no internal parameter to tune.

Table F2 – Probability thresholds selected.

Model	*Policy*	*Probability threshold*
Unpenalized logistic regression (LR)	A	0.2425
	B	0.5479
Penalized logistic regression (PLR)	A	0.24
	B	0.5424
Random forest (RF)	A	0.285
	B	0.5374
Gradient boosting machine (GB)	A	0.2675
	B	0.5453

Appendix G. Additional results

Table G1 – Kappa statistic of models in the restrictive test subset

	Kappa
Q-S25	0.045
Q-S50	0.067
Q-G	0.035
Policy A	
LR	0.172
PLR	0.182
RF	0.236
GB	0.248
Policy B	
LR	0.205
PLR	0.205
RF	0.287
GB	0.238

Note: The Kappa statistic discounts the amount of accuracy generated just by chance. Note that the chance-corrected accuracy of Q-S50 is low ($\kappa_{\text{QS50}} = 0.067$) and represents less than one third of the chance-corrected accuracy we could get with the random forest.

Table G2 – False alarm rates in the restricted test subset.

	FAR
Q-S25	0.054
Q-S50	0.767
Q-G	0.173
Policy A	
LR	0.531
PLR	0.526
RF	0.576
GB	0.452

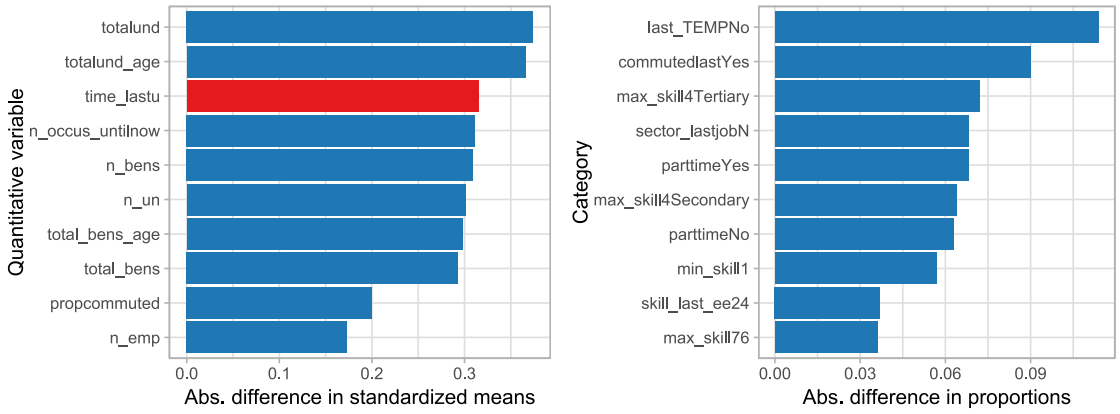
Note: FAR = 1 – specificity.

Table G3 – False alarm rates in two strata of the test subset.

	FAR
Older jobseekers	
LR	0.514
PLR	0.522
RF	0.469
GB	0.374
Female and older	
LR	0.531
PLR	0.541
RF	0.454
GB	0.367

Note: FAR = 1 – specificity.

Figure G1 – Top-10 differences in standardized means (left) or proportions (right) between the test subset and the restrictive test subset.



Source: Own elaboration. Denoting with μ_g a summary statistic for the dataset g, it is shown the difference $\mu_{test} - \mu_{restrictive}$. Therefore, blue bars denote a positive difference, whereas red bars collect a negative difference.

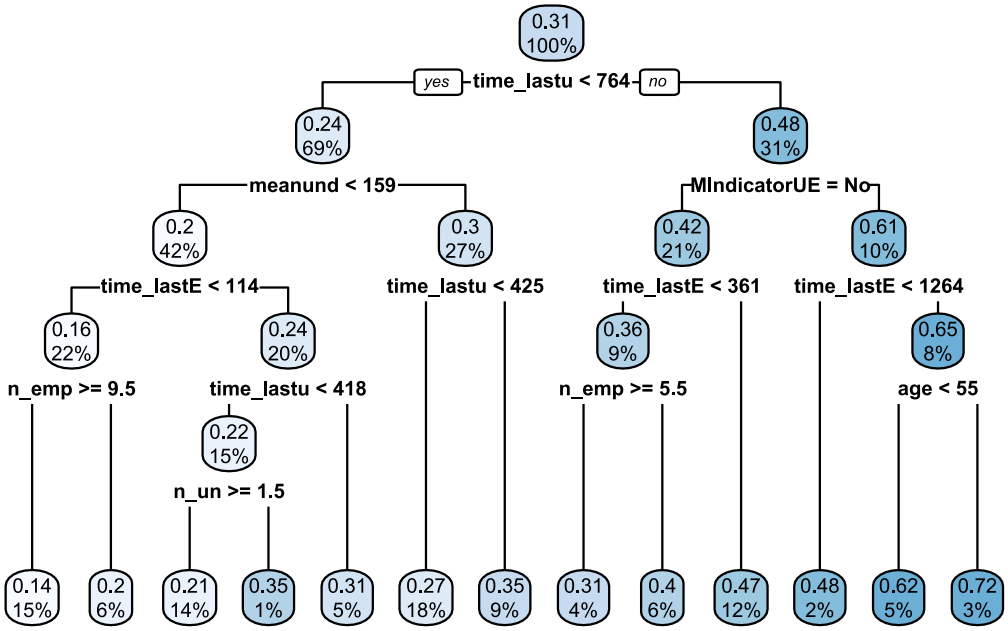
Appendix H. Global surrogate model

As an additional tool to interpretate how our statistical models make predictions, we have estimated a global surrogate model. The following regression tree model tries to forecast the predictions of the random forest model using 80 % of the training sample. The tuning parameters were fixed at the following values: the cost-complexity parameter equalled 0.005, the tree depth was 30, and the minimal node size was established at 2. The resulting model has a $R^2_{training} = 0.778$ and a $R^2_{test} = 0.774$, attaining a good approximation to the random forest with a relatively low interaction depth.

As shown in Figure H1, the tree incorporates seven covariates: the number of days since last unemployment episode (time_lastu), the number of days since the beginning of the last employment episode (time_lastE), the mean duration of unemployment episodes until the present (meanund), the number of employment episodes in the past (n_emp), the number of unemployment episodes in the past (n_un), the age (age), and the indicator of missingness on unemployment episodes in the past (MIndicatorUE). Note that all the predictors selected by the global surrogate model were also highlighted as remarkably important by the permutation-based variable importance statistic.

To interpret Figure H1, we must consider that each node shows the probability of experience a long-term unemployment episode and below the percentage of the sample that fits in each partition. Starting the partition from above, we see that the combinations of value that predict LTU with probability equal to 0.72 is: having the last unemployment episodes at least 764 days ago, having the last employment episode at least 1,264 days ago, and being older than 55 years. This profile is in line with the literature and fits with 3 % of our sample.

Figure H1 – Graphical representation of the decision tree



Chapter IV.

Employment effects of job search interventions in Veneto: assistance and intermediation[89]

1. Introduction

In the systematic reviews on the effects of ALMPs, job search programs are typically considered effective in increasing the employment probability (Vooren et al., 2019; Card et al., 2018). Looking deeper into the evaluations, it is difficult to draw lessons for policy designs as there are substantial differences in the characteristics of the intervention, the participants, or the evaluation methods. The variations on the details of the intervention have already been emphasized by many researchers asking for dissecting each program design during evaluations (OECD, 2006; Cronert, 2023). On one side, there is research indicating that job search assistance and job search intermediation have different effects because they act through different mechanisms (Cheung et al., forthcoming). This calls on researchers to clearly differentiate between the two channels. On the other side, evaluations usually focus on one program, which precludes comparing interventions with different intensities implemented on the same population. However, the treatment intensity or amount spent on each participant is a crucial factor to optimize in order to attain an efficient expenditure of the public budget.

In this chapter, we analyse the employment effects of two treatments with different doses of job search interventions. The program was implemented in the region of Veneto (Italy) assigning different endowments of (1) job search assistance through vouchers and (2) job search intermediation through performance pays to providers. The chapter tries to answer two questions. Firstly, what is the impact of different treatment endowments on the number of worked days. Secondly, what is the impact of these endowments on the quality of employment, which is proxied through the type of labour contract. Since we know that treatment endowments were assigned through a deterministic function, the method to answer credibly

[89] This chapter is based on a working paper titled "More money, more effect? Employment effects of job search programs in Veneto", which is authored by me.

our causal questions was the sharp regression discontinuity analysis. The identification of the treatment effect requires stronger assumptions, but estimating the effect of the treatment endowment is possible with weaker assumptions. For public finances, the effect of the endowment is important because it regards the amount of money reserved for a certain intervention during a certain period, which obviously has an opportunity cost[90].

The main findings of our study were the following. Regarding the duration of employment, the impact on the average of worked days was 0 for all semesters in the intermediate-low treatment comparison and of 10 days for the first year in the high-intermediate treatment contrast. The effect on the median of worked days is approximately one month for semester 3 in the intermediate-low comparison and for semester 2 in the high-intermediate contrast. The downside is that all these impacts disappear at the following semesters, which casts doubts on the value of the intervention. Concerning employment quality, the unique effect was an increase in the probability of having a short temporary contract as the longest episode in the two years after program began. This impact was only detected for the high-intermediate comparison. Thereby, the effects of increasing the endowment after two years are practically null in both dimensions of employment. After exploring plausible mechanisms, we found that participants in the focal treatment only spent one hour more in assistance instead of the 5 or 14 extra hours that allowed their endowment. This drives us to interpret the few effects detected as generated by the changes in the performance pays of job search intermediation.

This research makes two contributions to the literature. Firstly, to the best of our knowledge, this is the first work analysing the effect of different intensities of job search services in Europe. Although the large sample of experimental studies in USA allows a meta-analysis to estimate the impact of the intensity of the interventions, the number of evaluations of programs in Europe with a known assignment mechanism is quite low (Card et al., 2018). Secondly, we advance the literature on job search services implemented through vouchers. Unlike previous studies for Italy (Linfante et al. 2019; Scarano, 2023) or Germany (Ayaita et al., 2022; Winterhager et al., 2006), our study combines a credible identification

[90] This intention-to-treat approach is common in the literature of voucher effects (Doerr et al., 2016; Huber et al., 2018). In sum, it amounts to think the structure as a randomized treatment assignment with possible imperfect compliance.

strategy with a large sample size that offers greater statistical power to make comparisons between treatment groups.

The chapter goes on reviewing the literature related to our research and describing the design of the program evaluated. Section 3 explains the data used for our study and the methods applied to analyse them. The findings of this work are shown in Section 4, including some analyses on the plausibility of the identification strategy and the robustness of our results to modelling options. To end up, Section 5 offers some conclusions.

2. Job search services

2.1. Literature review

The literature on job search services is large and has the problem that usually conflates under a miscellaneous concept the tools of job search assistance and intermediation (Vooren et al., 2019; Card et al., 2018). According to our research questions, our investigation speaks directly to two literatures: the effect of different intensities of assistance and the effect of different intensities in intermediation. By intensity we mean the public expenditure invested in each intervention by participant, which may pursue more hours of intervention (Escudero et al., 2019; Crépon & van den Berg, 2016) or simply more productivity during the hours of such intervention (Martin, 2015)[91]. Due to the design of the program, this chapter is also related to the literature on vouchers in ALMP.

The first research line is the employment effect of job search assistance in function of their intensity. Liu et al. (2014) performed a meta-analysis of assistance interventions and described a null relationship between intensity (duration) and effect size after controlling for some population and evaluation features. Their review has a great internal validity, since 32 out of 47 studies were randomized experiments. The problem is that 69 % of these articles used a sample from USA and only four of these studies came from an EU country (Netherlands, Finland, or Hungary). Martinson et al. (2019) and Dyke et al. (2006) also arrived at the

[91] This concept of intensity combines the meaning it has in the macro and micro literatures on ALMP. It helps to accommodate program designs that condition the payment on the number of hours of the service (the classic outsourcing contract) or on the performance of such service (our case for intermediation).

similar conclusion of null effects with USA experiments. Many years before, Meyer (1995) reviewed other six assistance experiments carried out in USA but concluded that the intensity had a positive effect on employment. As we can see, there is a lack of credible evidence comparing the employment effects of different versions of the same program in terms of intensity for Europe.

The second research line is the employment effect of job search intermediation according to different intensities. The evaluation of job search intermediation is scarce, so the testing of different types of intermediation is even stranger. Bollens & Cockx (2017) studied three formats of vacancy referral (human remote, human face-to-face, and automatic) in Belgium, joint with sanctioning, and found positive effects for the three of them. The rest of studies did not analyse different versions of intermediation. Positive results also appeared in Germany and Switzerland with any intermediation under sanctioning (van den Berg et al., 2019; Zuchuat, 2023). With an experiment in Sweden, Engström et al. (2012) reported that the thread of monitoring did increase the applications after intermediation, but it did not increase the employment probability. On the contrary, Fougère et al. (2009) detected with observational data that intermediation rose the employment probability in a non-monitored country like France. Cheung et al. (forthcoming) reports a positive intention-to-treat impact on the treated of a job search program in Sweden, showing that intermediation instead of assistance might be the explanatory mechanism.

We have not identified any analysis on the effects of different intermediation intensities, but we can leverage the literature on performance-based payments. In our program, the different intensities required for each treatment group are recognized with different recompensations. This is also a way that PES use to cope with the incomplete contracts problem widely studied in labour economics (Cahuc et al., 2014). To study the effect of these clauses, we must first consider that in Europe the proportion of performance-based pay varies significantly across programs (Crépon, 2018)[92]. Our closest study is Zanella & Salomone (2024), who studied performance-based payments for private-provided intermediation in the region of Trento (Italy). Zanella & Salomone (2024) used experimental data and found that the intervention decreased the unemployment probability for men in 5

[92] Our program attaches 42-52 % of the total payment to the output. In United Kingdom, this proportion has arrived until the 80 %, but in Belgium has been sometimes only the 20 % (Crépon, 2018).

p.p. compared with a non-treated group. With public-provided job search intermediation, Burgess et al. (2017) found that it had a positive employment effect in small offices, but not in large ones. Koning & Heinrich (2013) analysed the application of performance-based pay in adult training programs and estimated that it had a small positive effect on some groups (recipients of unemployment insurance) and null on others (like people with disability)[93]. More generally, Rehwald et al. (2017) looked at the impact of these performance pay in other policy fields and found that the evidence was mixed.

Lastly, this evaluation is also related to the literature on the voucher provision of ALMP. In our case, the voucher involves free selection of service provider and of dose of assistance. We cannot test the effect of any of these two elements because we do not have a comparison group with closed provider or closed dose. However, it is useful to consider this literature to make hypotheses on the possible mechanisms behind our effect estimates. For Germany, the estimates on the employment effects of adult training with voucher versus no treatment were negative or null until three or four years later (Huber et al., 2018; Doerr et al., 2017). In Netherlands, Brenninkmeijer and Blonk (2012) analysed experimental data and detected that treatments with closed dose had positive effects, but treatments with open dose vouchers had outcomes equivalent to the untreated condition. The results for USA are not more hopeful, since they present unclear conclusions (Barnow, 2009).

Italy has already implemented other ALMP via vouchers. At the state level, Linfante et al. (2019) evaluated a similar program called *Assegno di Ricollocazione* and estimated that the intention-to-treat effect on employment was zero. At the regional level, Colombo et al. (2015) and Scarano (2023) have also analysed comparable programs in other territories of Northern Italy. They use an identification based on selection on observables, so we try to advance the literature by offering new results for Veneto and with greater credibility. Germany also used vouchers to implement job search services. Using a conditional lagged dependent variable strategy, Winterhager et al. (2006) evaluated their impact on the employment probability in Western Germany and found positive effects. Ayaita

[93] Kluve et al. (2019) also studied its use in adult training and other ALMP for young people. They reported that the use of these tools was associated to larger effect sizes.

et al. (2022) also used a similar identification strategy to study the effect on employment quality comparing private and public providers using vouchers.

2.2. Design of the program

«Assegno per il Lavoro» (AXL) was a program funded by the PES of Veneto for the whole region that was implemented from July 2017 to May 2022. Its main goal was decreasing unemployment. The original version of the program targeted this objective at workers older than 35 years old that had at least one of the criteria of social vulnerability defined in Government of Veneto (2017a). However, nine months later, the target population was redefined as all unemployed workers older than 30 years old.

The tools to achieve such outcome were job search assistance, job search intermediation, adult training, and internships[94]. The key point was that the intensity of these actions was different according to the treatment group in which the individual was classified. Table 1 gathers the endowments of each treatment group, disaggregating it by type of intervention. There are two main dimensions of variation among groups. Firstly, the amount of time (hours) devoted to job search assistance increases from group A to B, and from B to C. Specifically, group C more than doubles the number of hours of assistance available in group B. Secondly, the amount aimed at job search intermediation also changes by group. Although the variations are not as abrupt as in assistance, note that from group B to group C the endowment increases in a 50 % in this concept. To adapt the chronogram to the different intensities, each treatment group has a different deadline to spend the money. All treatment groups receive the same amount for adult training.

[94] Only individuals in group C may use the performance-based amount to fund their internships stipend. We do not have data on the participation in internships, but we argue that its use must have been low. The reason is that both the provider and the individual face a remarkable risk of non-payment. If the provider grants the stipend to the jobseeker and then the firm does not offer her an employment contract longer than 6 months (or the worker changes the job before the 6 months), the provider will have to face the payment of the stipend. Note that, unlike with intermediation services, this payment involves a cost *additional* to the salary of caseworkers. If the provider anticipates this situation and wants to postpone the pay of the stipend until the end of the AXL participation, the jobseeker will have to carry out the internships activities without getting any money. Internships in Veneto are regulated in Government of Veneto (2017c).

Table 1 – Endowments and duration of AXL according to the treatment group[95]

	Group A	*Group B*	*Group C*
Total endowment	3,536 €	4,264 €	5,796 €
- Job search assistance	266 € (7 hours)	494 € (13 hours)	1,026 € (27 hours)
- Adult training	1,770 €	1,770 €	1,770 €
- Job search intermediation*	1,500 €	2,000 €	3,000 €
- Internships*	0 €	0 €	
Maximum duration	3 months (6 with susp.)	5 months (8 with susp.)	6 months (9 with susp.)

Source: Government of Veneto (2017a and 2018a). All are nominal quantities. (*) These rows are performance-based payments. Thereby, it collects the *maximum* payment that a provider can get, in this case due to attaining an open-ended contract. If the provider certifies the sign of a temporary contract with duration greater than 12 months, this amount is multiplied by 0.8. If such contract has a duration between the 6 and the 12 months, it is multiplied by 0.6.

The assignment of individuals into treatment groups was carried out with a jobseeker profiling model. It was a statistical profiling model that used logistic regression to predict the probability of very long-term employment, i.e. being employed 24 months later. The model produced a score from the next regression equation[96]:

$$S_i = 1/\left[1+\exp\left(-\left(\beta_0+\beta_1 sex_i+\beta_2 age_i+\beta_3 foreign_i+\beta_4 studies_i+\beta_5 disab_i \right.\right.\right.$$
$$\left.\left.\left.+\beta_6 terminationreason_i+\beta_7 group_i+\beta_8 sector_i\right)\right)\right]$$

The value of the scoring variable is determined by sociodemographic factors (sex, age, foreign nationality, level of studies, and having a disability) and by factors

[95] This is the scheme of endowments and durations for the reformed version of AXL, the one passed in April 2019. The previous version included two small variations. First, group B had 2 extra months of maximum duration and group C had 3 additional months. Second, all groups had 554 € less for adult training.

[96] Government of Veneto (2017b) reports more information on the profiling model, including the estimates of the regression parameters.

on the last job (termination reason, major group of occupation, and sector of the company). We reversed S_i generating S_i in order to make the scoring and the treatment values follow the same direction (the larger, the more intense treatment). After being profiled, the individual is allocated into a treatment group following this decision function:

$$Group_i = \begin{cases} \text{A} & 0 \leq s \leq 0.4 \\ \text{B} & 0.41 \leq s \leq 0.60 \\ \text{C} & 0.61 \leq s \leq 1 \end{cases}$$

Once the jobseeker is inside a treatment group, she will receive a voucher that includes a proportion that will be paid after receiving the service and another proportion that will be paid based on the results after such service. With the voucher, the worker firstly chooses *which* private provider is going to provide her the services. Secondly, the worker and the caseworker agree on *how many* sessions of assistance to provide and on *which* adult training to undertake (if any). Note that the intermediation and the internships (if any) are only paid if the jobseeker signs a labour contract that lasted at least 6 months after receiving the service.

A typical warning on the unintended consequences of vouchers mixed with performance-based payment is cream-skimming, that is, the provider's behaviour of not selecting those individuals with the lowest employment probability (Koning & Heinrich, 2013). The design of the payment scheme of AXL may be understood as a recognition of the higher risk of not getting the payment after accepting an individual from group C by offering to providers a larger incentive to put effort of them. This design might avoid the creaming problem. Although it is not the focus of our study, this is confirmed if we analyse the local effect of being profiled B instead of A or C instead of B on the probability of participating in each provider. The results in Appendix A show that there is only evidence of cream-skimming for one provider, which only represents 1.52 % of the whole sample of participants.

3. Methods

3.1. Data

To answer our research questions, we have used administrative microdata provided by the PES of Veneto (*Veneto Lavoro*). The dataset collects the universe

of individuals that wanted to participate in AXL at any moment of its implementation period.

The outcome variables were three, although they were observed at different times. To study the employment duration, we have used as outcome the number of days worked as a salaried employee at the first, the second, the third, or the fourth semester since the treatment began. To analyse the employment quality, we have utilized as response a categorical variable measuring the type of labour contract attached to the longest employment episode registered during the four semesters since treatment started. Note that in the definition of employment quality proposed by Muñoz de Bustillo et al. (2011), one of the two dimensions is the type of labour contract. This variable distinguishes between five types of contract: open-ended, fixed-term longer than 12 months, fixed-term between 6 and 12 months, fixed-term between 0 and 5 months, and no employment[97]. This variable gathers the expected (declared) duration of the contract, not the actual duration. To explore possible mechanisms, we have collected a third set of outcomes indicating the number of treatment doses taken by the participant. We know the number of hours of assistance received by each jobseeker and also the number of hours of adult training agreed by each participant[98].

The scoring variable is the unique confounding factor by design, but we have information on other variables as well. We received data on sociodemographic covariates that will be used for the plausibility analyses and for heterogeneity analyses acting as possible effect modifiers. A complete list of variables with more detailed definitions is available in Appendix B and a table with summary statistics of our sample is included in Appendix C.

3.2. Analysis

3.2.1. Target parameters

To accommodate several outcomes, Y_{oi} denotes the dependent variable for outcome o and individual i. However, we are going to eliminate subindex o to reduce the notational burden. Starting from the potential-outcomes framework

[97] The reason to categorize the contracts in this way is that performance-based payment is based on these categories. Only the last two categories do not receive a payment.

[98] Due to availability reasons of the data provider, we do not have information on the actual number of hours of adult training received.

with binary treatments (Imbens & Rubin, 2015), the potential outcome after participating in the active treatment is denoted as $Y_i(1)$ and the potential outcome after participating in the control treatment is denoted as $Y_i(0)$.

There are two active treatments whose effect we want to estimate, so we will use as their control treatments the group with prior intensity. To carry out each treatment comparison independently, we split the sample in those who are either in group B or group A (for the contrast D_{1i}) and in those who are either in group C or group B (for the contrast D_{2i}). Firstly, we define D_{1i} as a binary variable registering if individual i was treated in group B (the active treatment, taking value 1) or in group A (the control treatment, taking value 0). Secondly, we define D_{2i} as another binary variable collecting if individual i was treated in group C (the active treatment, taking value 1) or in group B (the control treatment, taking value 0)[99]. Both are functions of $Group_i$, which in turn is fully determined by the running variable S_i. Lastly, we recentre the running variable in each sample locating the threshold at $c = 0$ by generating S_{1i} for the first sample and S_{2i} for the second. Thereby, we have D_{pi} and S_{pi} for $p \in \{1,2\}$. Without loss of generality, we remove subindex p just for clarity of exposition.

We are interested in three target parameters: local average treatment effect of D_i on Y_i, denoted as $LATE$; local average treatment effect of D_i on the probability of $Y_i = j$, abbreviated as $LATE_j$; and local quantile treatment effect for a certain quantile $\mathbb{Q}_\tau$ at percentile τ of D_i on Y_i, denoted as $LQTE(\tau)$. In an equivalent way, we may think them as local effects of the (active) endowment or as intention-to-treat effects of the (active) treatment. They are formally defined as follows for each treatment comparison p:

$$LATE = \mathbb{E}[Y_i(1) \,|S_{pi} = c] - \mathbb{E}[Y_i(0) \,|S_{pi} = c] \text{ for } p \in \{1,2\}$$

$$LATE_j = \Pr[Y_i(1) = j \,|S_{pi} = c] - \Pr[Y_i(0) = j \,|S_{pi} = c] \text{ for } p \in \{1,2\}$$

$$LQTE(\tau) = \mathbb{Q}_\tau[Y_i(1) \,|S_{pi} = c] - \mathbb{Q}_\tau[Y_i(0) \,|S_{pi} = c] \text{ for } p \in \{1,2\}$$

3.2.2. Identification

The target parameters mentioned in the past section include counterfactual potential outcomes, so we need identification assumptions to estimate them with

[99] We may equivalently think that we are estimating two treatment comparisons.

the observed data. Following Rubin (2007; 2008), we argue that an identification strategy is more credible the closer it aligns with the properties of a randomized experiment[100]. Due to the program design of AXL, the most credible identification strategy is the sharp regression discontinuity. Let us present the identification assumptions in relation to the assignment mechanism of a completely randomized experiment.

The first identification assumption deals with the unconfoundedness of the assignment mechanism. It states that once we condition on the scoring variable, the assignment does not depend on any potential outcome. This is a weak assumption in regression discontinuity designs, since they are based on detailed knowledge of the functional form of the treatment assignment. Notice that this assumption equals the property of randomized experiments that posits that the assignment mechanism is fully known apart from unconfounded. This first identification assumption may be formalized as follows (Abadie & Cattaneo, 2018; Frölich & Sperlich, 2019):

Assumption 1. The assignment mechanism is $D_i = \mathbf{1}(S_i \geq \mathrm{c})$, so $\lim_{s \downarrow \mathrm{c}} \mathbb{E}[D_i|S_i = s] - \lim_{s \uparrow \mathrm{c}} \mathbb{E}[D_i|S_i = s] \neq 0$.

Assumption 1 would be violated if the treatment was not only a function of the score, but also a function of unobserved variables like the ability of the jobseeker or her personal network. The literature has called this concern the score manipulation problem to highlight that these unobservables might alter the score (Frölich & Sperlich, 2019). On the one hand, we will analyse each ingredient of the scoring variable to judge how feasible is to manipulate its value. On the other hand, we will carry out the density tests proposed in Cattaneo et al. (2020). Like pre-treatment trends in difference-in-differences analyses, passing this test is neither sufficient nor necessary for Assumption 1 to hold. Still, it is usually used by researchers to evaluate the credibility of the identification.

The second identification assumption faces the lack of overlap caused by a deterministic assignment mechanism. On the contrary, an important property of

[100] Or, equivalently, the less it deviates from it.

a randomized experiment is a probabilistic assignment mechanism, which is sometimes simply called overlap. Overlap ($0 < \Pr[D_i = d|S_i = s] < 1$) is absent in our case due to the program design, which assigns treatments deterministically. For instance, we know that if individual i has $S_i > \mathrm{c}$, then $\Pr[D_i = 0|S_i > \mathrm{c}] = 0$. Cattaneo et al. (2019) presents two approaches to deal with this problem. The first is the local randomization approach and requires having a natural experiment in the sense of Titiunik (2021). Due to the lack of chance in our treatment assignment mechanism, we argue that this approach does not fit with our setup. The second is the continuity approach, which involves making a small extrapolation around the threshold. The key idea is that extrapolations that are very far from the observed data are more model-dependent (King & Zeng, 2006), so this approach minimizes such risk by only focusing on the threshold of treatment assignments. This is done by modelling the expectation of the treated potential outcome function and the expectation of the untreated potential outcome function as continuous functions. For our three target parameters, this assumption may be written as follows (Frölich & Sperlich, 2019):

Assumption 2.

- For $LATE$: $\mathbb{E}[Y_i(1)|S_i]$ is continuous at $S_i = \mathrm{c}$ and $\mathbb{E}[Y_i(0)|S_i]$ is continuous at $S_i = \mathrm{c}$.
- For $LATE_j$: $\Pr[Y_i(1) = j\,|S_i]$ is continuous at $S_i = \mathrm{c}$ and $\Pr[Y_i(0) = j\,|S_i]$ is continuous at $S_i = \mathrm{c}$.
- For $LQTE(\tau)$: $F_{Y(1)|S}(y|\mathrm{c})$ is continuous and $F_{Y(0)|S}(y|\mathrm{c})$ is continuous.

Two plausibility analyses may enhance the credibility of this assumption. Firstly, one threat to this condition is that other intervention used the same function to assign treatments. In this case, $\mathbb{E}[Y_i(0)|S_i]$ might not be continuous and face a jump just at the cutoff if there was an alternative treatment with non-null effect. To assess the plausibility of Assumption 2, one side, we perform placebo outcome tests just at the cutoff (Eggers et al., 2024). This exercise is also known as pseudo-treatment analysis (Frölich & Sperlich, 2019), since it looks at outcome differences in a period in which the individual had not yet participated in AXL. Evidence on statistically significant differences would cast doubt on Assumption 2. If the

intervention was assigned as designed, it would indicate the presence of an alternative treatment in the past with the same assignment mechanism. Of course, this does not involve that the same action will be present in our periods of interest. Following Manski & Pepper (2018), we take deviations in the past as a guide to estimate the size of the bias in the focal periods. We will also look at the qualitative data (texts) of AXL to speculate on the possibility of alternative programs using the same scoring function. Secondly, another threat to the continuity of the untreated potential outcome functions might come from differences between treatment groups in key covariates. According to Assumption 2, once we control for the score, these functions should not make jumps. Having treatment groups balanced in variables that might influence the outcome neutralizes the confounding problem, so covariate balance will be also assessed.

3.2.3. Estimation and inference

Following the convention in regression discontinuity analyses, we use visual and numeric techniques to present the results of the estimates. Firstly, we employ regression discontinuity plots following the binscatter procedures proposed by Cattaneo et al. (2024). These figures join scatterplots of local means, proportions, or medians with global regression curves for each side of the cutoff. Secondly, we present quantitative estimates with more detailed estimation and inference methods.

With respect to point estimation, we employ one estimator for each estimand. For the $LATE$, we apply a non-parametric method: the local polynomial estimator (Cattaneo & Titiunik, 2022). The bandwidth is common for both sides of the threshold and will be selected with the minimization of the mean squared error (Calonico et al., 2017). The degree of the polynomial estimator will be chosen through the procedure of Pei et al. (2022), who focuses on the minimization of the asymptotic mean squared error. We will try both linear and quadratic specifications in the score, always with a triangular kernel for consistency with the rest of analyses[101]. For the $LQTE(\tau)$, we use the local estimator described in

[101] Moreover, the triangular kernel has the property of point estimation optimality and is compatible with the technique of Pei et al. (2022). Both the triangular and the Epanechnikov kernel are not uniform, i.e. give more weight to observations near the cutoff. The results are robust to using a uniform kernel for almost all time horizons and are available upon request.

Qu & Yoon (2019). The bandwidth selection follows the same idea as before, i.e. minimizing the mean squared error. Three alternative procedures based on this intuition are executed and, to avoid bias, we choose the one that offered the smallest bandwidth[102]. We follow Qu & Yoon (2019) and use a linear specification with Epanechnikov kernel. To avoid ungrounded assumptions, quantile regressions will not be overall interpreted using the rank invariance assumption (Dong & Shen, 2018). Therefore, estimates of $LQTE(\tau)$ do not provide evidence on who are the individuals most affected by each treatment. However, they are informative on changes in the distributions of the potential outcomes, which aids in assessing the overall value of the program.

For the $LATE_j$, we apply the maximum likelihood estimator proposed by Xu (2017). Departing from a multinomial logit model, we place the level of "not employment" as the reference category. The bandwidth takes again the mean squared error as criterion, but with its necessary adaptation to the categorical context. The procedure recommended by Xu (2017) chooses a bandwidth jointly for all the non-reference categories of the outcome. We employ a linear degree for the parameters of S_i and a uniform kernel, since the findings of Xu (2017) discourage a non-uniform kernel.

With respect to inference, the methodological literature has tried to remove the bias that remains after selecting a bandwidth by minimizing the mean squared error, since this is a mix of bias and variance (Cattaneo et al., 2019). Although all are Wald tests, methodologists have developed different inference procedures for each estimand. For the $LATE$, we apply a test of null effect using the robust-bias corrected proposal of Calonico et al. (2014). For the $LQTE(\tau)$, we employ two hypothesis tests developed by Qu & Yoon (2019), which are robust bias-corrected as well. Firstly, we want to investigate if all treatments had a null effect on the selected quantiles. Therefore, the hypotheses are $\mathrm{H}_0\colon LQTE(\tau) = 0$ for all $\tau \in \mathcal{T}$, denoting with $\mathcal{T}$ the quantile range, and $\mathrm{H}_1\colon LQTE(\tau) \neq 0$ for some $\tau \in \mathcal{T}$. Secondly, we want to know whether the effect of the intervention was the same at all quantiles of interest. Thereby, the hypotheses are $\mathrm{H}_0\colon LQTE(\tau)$ is constant over $\tau \in \mathcal{T}$, with $\mathrm{H}_1\colon LQTE(\tau) \neq LQTE(\upsilon)$ for some $\tau, \upsilon \in \mathcal{T}$. For the $LATE_j$, we have a test of zero effect

[102] We tried the bandwidth selector of Imbens & Kalyanaraman (2012) and the two selectors of Qu & Yoon (2015), one that treated the cutoff as interior point and another that considered it a boundary point.

proposed by Xu (2017) that translates the robust bias correction of Calonico et al. (2014) to the categorical context.

We will assess the robustness of our findings to changes in small modelling decisions like the degree of the polynomial, perturbations of the bandwidth, or different bandwidths for each side of the cutoff. Lastly, we will also carry out a heterogeneity study in the form of subgroup analyses. This method simply stratifies the sample according to the values of one variable that is believed to be a possible effect modifier. Then, the same analytical strategy used in the main analysis is applied on the subpopulation retained in each strata.

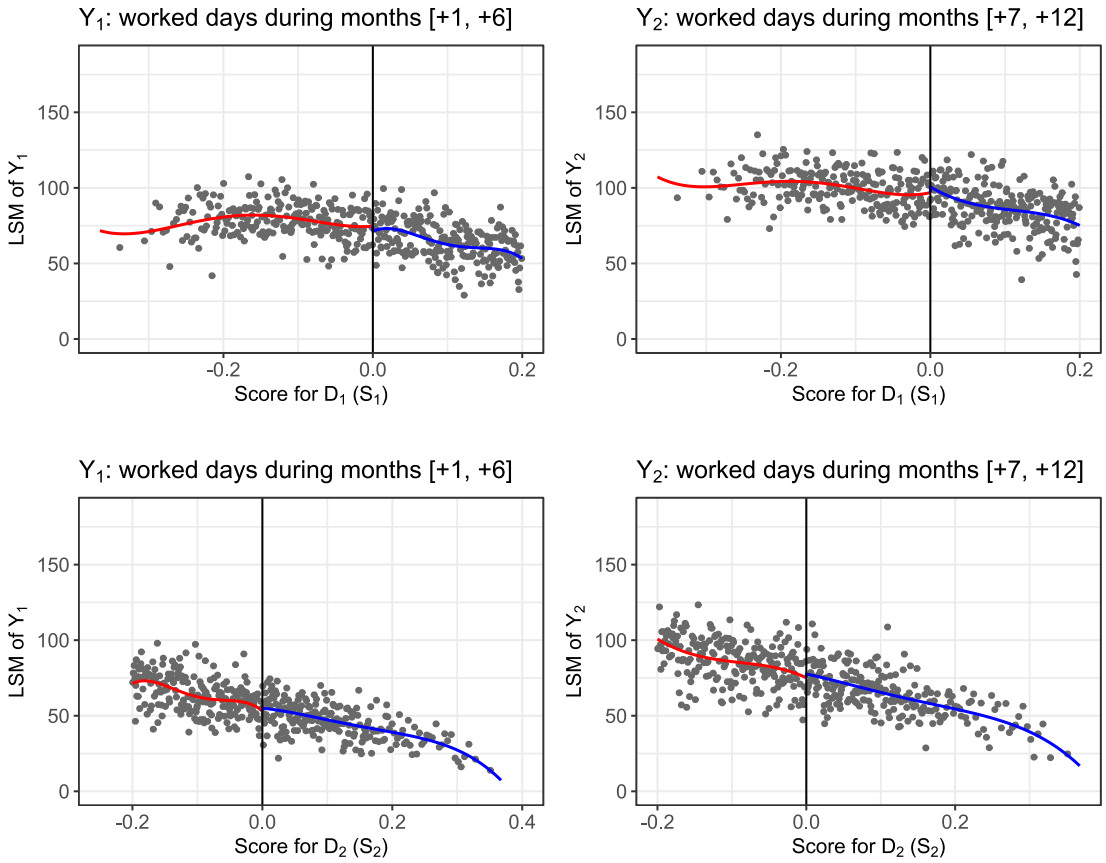
4. Results

In this section, we present the estimates on the employment effects of two comparisons of treatments offered under the AXL program. Then, we report the findings on the effects of treatment assignment on the actual number of treatment hours received as a possible mechanism. Lastly, we offer some plausibility and robustness checks to enhance the credibility on the internal validity of our results.

4.1. Employment effects

Let us start with the visual evidence about the local effects on the employment duration. Figure 1 reports regression discontinuity plots including local means on the number of worked days for both treatment comparisons (D_1 and D_2) and the first and second semesters as horizons (Y_1 and Y_2). The plots for the next semesters are provided in Appendix D. Remind that D_1 contrasts the groups with intermediate and low endowments (B vs. A), while D_2 does the same for the high and intermediate endowments (C vs. B). The main finding from these graphs is the absence of a clear jump just at the cutoff, which would indicate a zero local effect both for the first and the second rises in the treatment endowments. Nonetheless, these figures need to be complemented with more quantitative evaluations of the program.

Figure 1 – RD plots with local means for employment duration variables (semesters 1 and 2)



Note: Regression discontinuity plots for semesters 3 and 4 can be found in Appendix D. LSM: local sample mean.

The previous plots convey a clear message but may not be totally accurate due to the use of global parametric curves at each side of the cutoff. To offer a more flexible analysis, Table 2 shows local estimates of the $LATE$. The first part reports the average effect of D_1 just at the threshold for the four semesters. The coefficients are not statistically significant, which may be interpreted as null effects. Beyond statistical inference, the point estimates range from -2 to 8 worked days, which is a small amount considering the number of days in a semester. The second part displays the average effect of D_2 at the cutoff. If we departed from the hypothesis of a linear dose-response function between the endowments and the worked days, the changes estimated in this case should be larger due to the bigger difference of endowments of D_2. The effect estimate is only statistically significant at the usual $\alpha = 0.05$ level in the second semester, attaining a modest

effect size of 10 days for this period. The following semesters do not show a statistically significant difference between groups at the threshold. Thereby, at least for those at the cutoff, we do not have evidence supporting the idea of a positive effect of the two endowments increases in almost any semester.

Table 2 – Estimates of employment duration effects for each treatment at the mean (semesters 1 to 4)

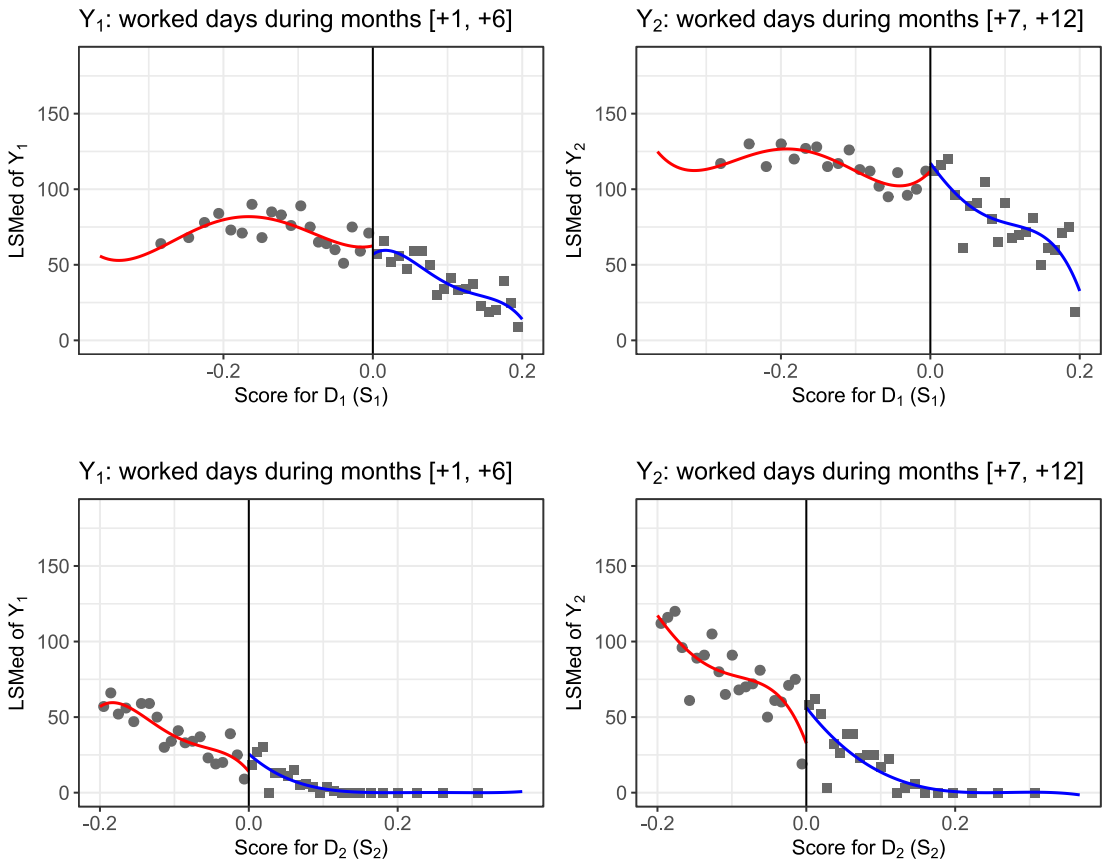
	Y_1	Y_2	Y_3	Y_4
D_1	-1.745	4.837	8.075	4.587
	(3.715)	(4.174)	(5.115)	(4.738)
Bandwidth	0.063	0.062	0.045	0.053
N_h	6564	6400	4719	5507
D_2	6.835*	10.378**	4.201	7.512
	(4.387)	(4.929)	(4.806)	(5.341)
Bandwidth	0.037	0.043	0.048	0.043
N_h	4079	4741	5246	4640

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell offers the point estimate of the $LATE$ and its standard error in parenthesis. N_h is the effective sample size. The polynomial degree is always 1.

The previous results show null (or negligible) effects of both endowment increases at the mean, but the story might be different if we looked at the quantiles of the distribution of worked days. Figure 2 attends this possibility by plotting local medians on the worked days for treatment comparisons D_1 and D_2 during the first and the second semesters. Appendix D includes the rest of plots. On one side, the plots of the first active treatment (D_1) do not reflect a clear jump in the number of worked days at the threshold. On the other side, the plots of the second active treatment (D_2) do represent a different situation depending on the

semester. In fact, the functions for semester two show a larger difference at the cutoff. This invites us to carry out a closer examination with more fine-grained techniques.

Figure 2 – RD plots with local medians for employment duration variables (semesters 1 and 2)



Note: RD plots for the number of worked days at the third and fourth semesters are available in Appendix D. LSMed: local sample median.

The results in Table 3 illustrate a more detailed approach on the $LQTE(\tau)$, because they offer estimates on all the quartiles and because they employ non-parametric regressions. In general, the results in the table are in line with the findings for the two first semesters shown in the plots. The first panel of results deal with the effects of the active treatment in D_1, i.e. group B vs. group A. Firstly, it is noteworthy that the point estimates are equal to zero for the first quartile in all semesters. Note that, under the rank invariance assumption, this is

an area of the distribution where the most vulnerable workers are located. Secondly, the point estimate on the effect at the median shows an increase of around one month of employment in semesters three and four. However, in terms of statistical inference, we may only discard the hypothesis of null effect at any quartile for semester three. For semester three, there is also evidence of heterogeneous effects between quartiles. Thereby, we conclude that the active treatment had a positive effect at the median of the distribution of worked days[103]. The bad news is that it seems that this effect vanishes, since we do not have statistically significant evidence on a non-zero impact at the next semester.

The second panel of results in Table 3 shows the effects of the active treatment in D_2, i.e. group C vs. group B. The first finding is a positive effect at the median during the first year, attaining an extra month of employment at the median of the distribution during the second semester. During these semesters, we can reject the hypothesis of null effect in all quartiles on the number of worked days. However, these impacts disappear at the following semesters. The second finding is that the changes at the first and third quartiles are almost non-existent in any semester. This make us think that the rejections of the homogeneity test during the first and second semester are actually collecting an effect at the median. Compared with the effects at the average, we find that D_2 in semester 2 had a larger effect at the median than at the mean, which fits with a long tail of no changes in the distribution at the left of the median[104].

[103] At first sight, it might seem difficult to reconcile an effect at the median larger than at the mean when (1) we cannot reject that the mean effect was zero and (2) the rest of quartiles have non-negative point estimates. Following the convention in the econometrics literature, our point estimators use a bandwidth that minimizes the sum of the squared bias *and* variance. This involves that some bias might remain at the benefit of a higher precision. If we use a bias-corrected point estimator, the point estimates of quartiles 1 and 3 are negative but not statistically significant. This decrease in bias comes at the cost of a decrease in precision. We argue that this negative effects at the other quartiles would explain this bigger effect at the median than at the mean. These robustness analyses are available upon request.

[104] If we perform an analysis with larger resolution, i.e. estimating effects at the deciles, we see that D_2 had no effect at deciles 1, 2, 3 and 4. These analyses are available upon request. Anyway, in discrete variables, it is possible than the true mean was lower than the true median while having a long tail at the right-hand side (see Von Hippel, 2005).

Table 3 – Estimates of employment duration effects for each treatment at the median (semesters 1 to 4)

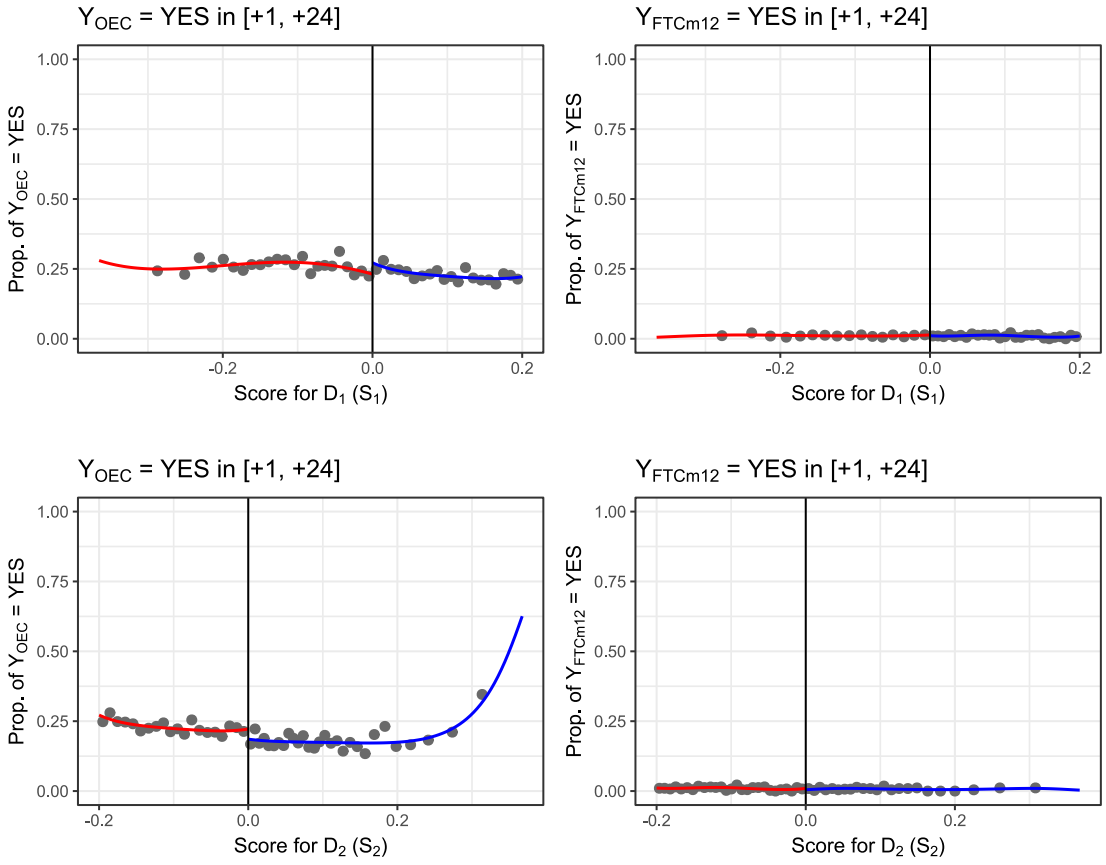
	Y_1	Y_2	Y_3	Y_4
D_1				
$\tau = 0.25$	0	0	0	0
$\tau = 0.5$	-4.85	10.54	27.638	25.486
$\tau = 0.75$	2.862	1.015	0.077	1.089
Treatment non-significance test	.	.	**	.
Treatment homogeneity test	.	.	**	.
D_2				
$\tau = 0.25$	0	0	0	0
$\tau = 0.5$	5.775	29.944	-6.434	0.023
$\tau = 0.75$	2.998	2.407	0	0
Treatment non-significance test	***	***	.	.
Treatment homogeneity test	***	***	.	*

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The first three rows of each panel report the point estimate of the $LQTE(\tau)$.

So far, we have seen that AXL had not lasting effects on the employment duration. However, it might be the case that it had on the employment quality. Although this variable does not allow a disaggregation by semesters, Figure 3 shows regression discontinuity plots on the probability of signing each type of labour contract during the two years since treatment began. To be concise, we only

include graphs for the two contracts with highest employment quality, i.e. the open-ended (OEC) and the long fixed-term contracts (FTCm12). The rest of plots are available in Appendix D. The plots for the OEC outcome in both treatment comparisons presents a small discontinuity, which might indicate an effect of the endowments. The plots for the FTCm12 do not barely show any jump, so it seems there was no effect on the probability of having this contract as the longest one.

Figure 3 – RD plots with local proportions for employment quality variables collecting if the longest contract was open-ended (OEC) or temporary for more than 12 months (FTCm12)



Note: RD plots for the probability of signing a fixed-term contract of [6, 12] months and the probability of signing a fixed-term contract of [0, 5] months are in Appendix D.

In order to have a closer examination of the effects on employment quality, we estimate the differences between groups non-parametrically and present the

results in Table 4. This table includes estimates not only for the contracts with more stability, but also for temporary contracts of intermediate (from 6 to 12 months, denoted as FTC612) and short duration (from 0 to 5 months, denoted as FTC05). For both treatment comparisons, the point estimates are in general not statistically significant at the usual level of $\alpha = 0.05$. The unique exceptions appear in the effects of the treatment endowment C versus B (D_2), for which we have evidence on a 4.5 p.p. increase in the probability of signing a short temporary contract (FTC05) as the longest contract. The increase in these short contracts has been accompanied by a decrease of 3.1. p.p. in the probability of obtaining a permanent contract (OEC). This is consistent with the results of the quantiles, where we saw an increment of around one month for D_2 at the median only for the second semester. Still, it is an intriguing result, since short temporary contracts are not sufficient to receive the performance-based payment of intermediation. A possible reason is that caseworkers expected that this short contract might work as a stepping-stone for a longer labour contract that could justify the intermediation payment. There is also other coefficient of modest size but statistically significant if we accepted a significance level of $\alpha = 0.1$. Focusing us on the most stable contract, the effect on the probability of signing an open-ended contract seems positive for D_1, whereas it appears negative for D_2.

Table 4 – Estimates of employment quality effects for each treatment and contract

	Y_{OEC}	Y_{FTCm12}	Y_{FTC612}	Y_{FTC05}
D_1	0.016*	0.000	0.018	-0.017
	(-0.004, 0.059)	(-0.009, 0.006)	(-0.011, 0.044)	(-0.059, 0.013)
D_2	-0.031**	0.001	-0.003	0.045***
	(-0.063, -0.004)	(-0.004, 0.008)	(-0.034, 0.019)	(0.019, 0.092)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate of the $LATE_j$ and the 95 % confidence interval in parentheses.

4.2. Mechanisms

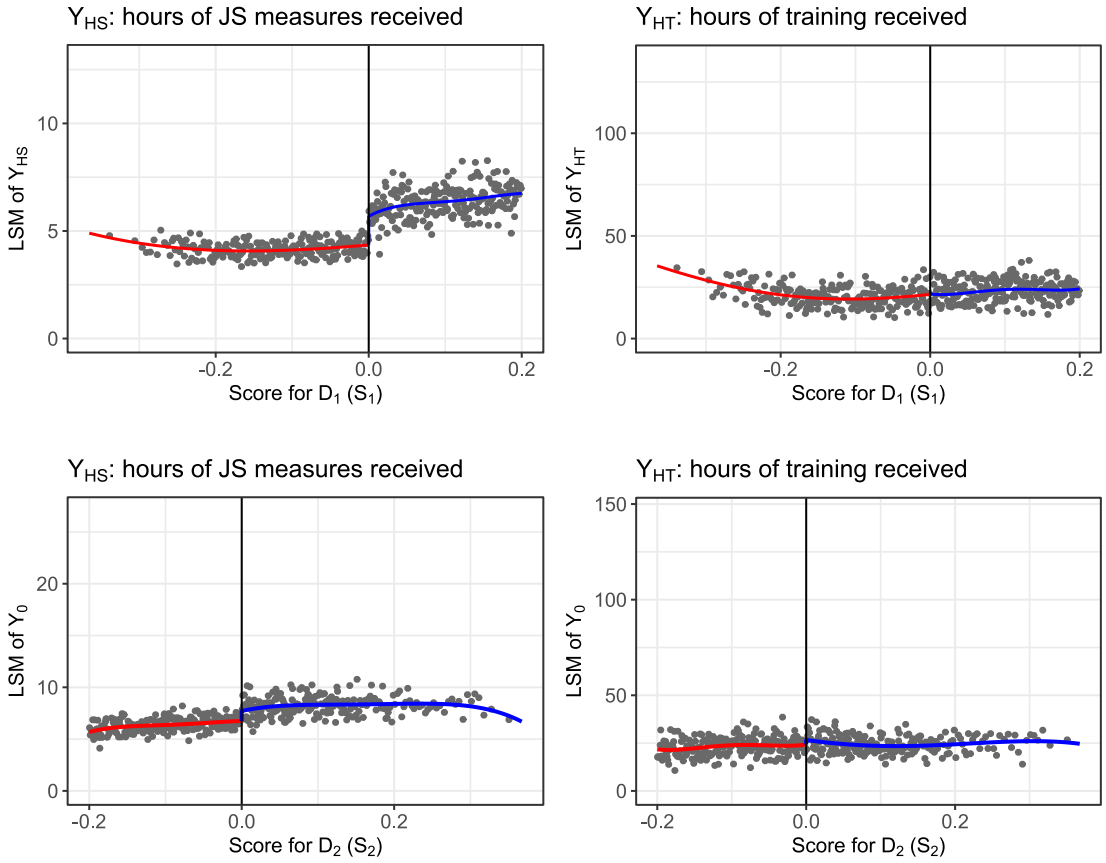
We have seen that the effects of AXL on the employment duration and the employment stability were overall null or almost null for the two treatment comparisons. The following question is to discern why the program did not achieve its objective. To put it differently, we want to know which mechanisms of the policy toolbox did not work as expected. The program design allows us to credibly identify one important mechanism: the treatment uptake or compliance. We maintain that it is reasonable to assume that policymakers provided a higher endowment to those with less employment probability to foster a more intense intervention. Thereby, a partial expenditure of the endowment may be understood as a failure of compliance.

The regression discontinuity plots of Figure 4 show the local means of hours received in assistance (Y_{HS}) and in adult training (Y_{HT}) for the two treatment comparisons. Regarding the first contrast (D_1), we see that there is a jump just at the threshold in the hours of assistance. On the contrary, we do not find evidence of a discontinuity in the case of hours of adult training. Regarding the second contrast (D_2), the hours invested in assistance seem to vary slightly just at the cutoff[105]. Again, the situation of the adult training for D_2 looks invariant to the total endowment assigned to each group. In sum, despite the different assignments of each treatment group, the results on the uptake at the cutoff indicate that jobseekers have not substantially varied their uptake in function of their endowment. It seems reasonable to claim that this low compliance influenced the null effects estimated in the last section[106].

[105] The scales of the vertical axis in the plots about Y_{HS} are different for D_1 and for D_2. This has been intentional to indicate the differences in the maximum number of hours allowed by the largest endowment of each comparison.

[106] We can identify the effect of the assignment on the uptake with weak assumptions (those of the sharp regression discontinuity shown in Section 3.2). However, we require much stronger identification assumptions to identify the effect of the treatment *received* on the different outcomes. The reason is that it is remarkably difficult to propose a credible model for the assignment mechanism that would explain why people took different treatment doses.

Figure 4 – RD plots for the effect on treatment intensity



Source: Own elaboration. LSM: local sample mean.

To have a more flexible estimation of the effects of the assignment on the uptake, Table 5 shows the estimates using local regressions for each comparison and each outcome. In general, the findings are in line with the visual evidence. Regarding assistance, there is an average difference of one hour of uptake between active and control treatments both in D_1 and in D_2. The coefficient is statistically significant at $\alpha = 0.01$ in both comparisons. The size of the differences in assistance uptake is striking, since it is remarkably small if we take into account that the potential difference between treatment groups is of 6 hours (for D_1) and 14 hours (for D_2). Concerning adult training, there is an average difference in uptake of zero hours for both comparisons, since the coefficients are not statistically significant. All groups received the same amount for this action, so this result is less unexpected. In conclusion, taking into account the tiny differences in treatment uptake of assistance and adult training, the findings of Section 4.1 might be interpreted as the impact of different doses of intermediation.

Table 5 – Estimates of effects for each treatment assignment at the mean of hours of treatments receipt

	Y_{HS}	Y_{HT}
D_1	1.252***	-0.216
	(0.211)	(1.331)
Bandwidth	0.045	0.088
N_h	4685	9233
D_2	1.110***	2.209
	(0.256)	(1.724)
Bandwidth	0.074	0.055
N_h	7973	6011

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The cell shows the point estimate of the *LATE* and the standard error in parentheses. N_h is the effective sample size. The polynomial degree is always 1.

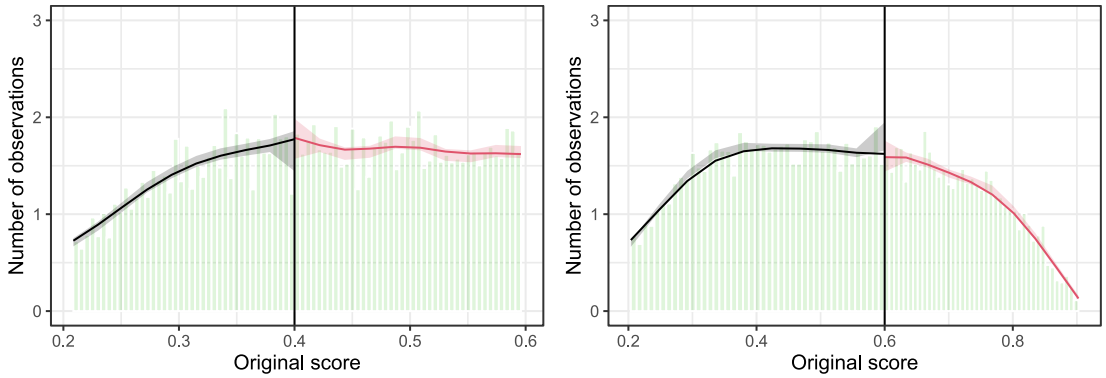
4.3. Plausibility and robustness

In this section, we are going to run analyses to get some clues on how plausible our identification assumptions are and to check how robust are our main findings to variations in the modelling options. The plausibility checks are the study of score manipulation, placebo outcome analyses, and covariate balance analyses. The robustness analyses deal with changes in the bandwidth, the kernel, the polynomial degree, or the population of interest.

The first plausibility check looks for differences at the cutoff in the probability density of the score. Being this a continuous variable, if the assignment mechanism actually implemented was the one posit in the first identification assumption, we should find equal densities close to the threshold. Figure 5 presents estimations of the density function with 95 % confidence intervals. We can see how the confidence intervals overlap in both cases, because we cannot reject the hypothesis

of continuity of $f_S(c)$ at the usual significance level for both cutoffs. The conventional interpretation is to take this as good news, since it supports the idea of an assignment mechanism governed by the running variable. Additionally, we also reviewed qualitative data on this scoring variable, i.e. documents of the program design. Our conclusion was that most ingredients of the score are either impossible (e.g. variable "age") or highly costly (e.g. variable "level of studies") to manipulate.

Figure 5 – Histograms and estimated densities for D_1 and D_2



Note: The left plot signals the cutoff for D_1, whereas the right plot signals the cut-off for D_2.

The second plausibility analysis uses placebo outcomes to get some insights on the possible simultaneous execution of another program with the same running variable. Using the staggered implementation of AXL, we can take individuals that had not yet participated at certain periods in which the program had already started[107]. Then, we can estimate the "local effect" of being included in the

[107] We are going to use just the semester prior to starting the treatment (Y_{-1}) to avoid using samples with different number of observations for each pre-treatment period. For instance, if we would like to estimate the pseudo-treatment effect at two semesters prior to starting the treatment (Y_{-2}), we would need to discard those individuals treated in semester 1 and in semester 2. The reason is that we established the requirement that these pseudo-treatment effects should be estimated when the program had already begun. If the program started in $t = 0$, getting those treated in $t = 2$ would involve observe their outcome in $t = -1$, when the program was not active yet. Using different sample sizes has the problem of departing from tests with different statistical power.

treatment groups created by D_1 and D_2. Detecting differences in these comparisons would question the continuity of the untreated potential outcome functions, which is key for the second identification assumption.

Table 6 – Estimates of employment duration effects for each pseudo-treatment

	$LATE_{Y_{-1}}$	$LQTE(0.5)_{Y_{-1}}$
D_1	-2.625 (3.546)	-13.144
P-value	0.569	TS: $0.7 > p > 0.6$ TH: $0.6 > p > 0.5$
D_2	-4.761 (3.524)	0
P-value	0.242	TS: $0.6 > p > 0.5$ TH: $0.3 > p > 0.2$

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Column 1 reports the point estimate of the $LATE$, its standard error and the p-value for the hypothesis of null effect. Column 2 presents the point estimate of the $LQTE(0.5)$ and the p-value of each hypothesis (TS: treatment non-significance test, TH: treatment heterogeneity test). The polynomial degree is always 1.

Table 6 shows the estimates of the effects of these pseudo-treatments on the average and on the median of worked days in the semester previous to starting AXL[108]. On the $LATE_{Y_{-1}}$, the coefficients of both comparisons are not statistically significant at the usual significance level, so this would indicate null differences. On the $LQTE\ 0.5\ _{Y_{-1}}$, the point estimate of the difference is around thirteen days for D_1, but we cannot rule out that this difference is due to sampling noise. The reason is that neither the hypothesis of treatment non-significance nor the hypothesis of treatment homogeneity may be rejected at the common levels. In

[108] We cannot carry out these placebo outcome analyses with the employment quality outcomes because we do not have data on the employment episode with the longest expected duration before the participation in AXL.

conclusion, the null evidence on the presence of differences in the groups before treatment participation supports the idea of the non-existence of alternative treatments with non-null effects and the comparability of treatment groups at the threshold.

The third plausibility analysis is the study of covariate balance, which checks how comparable are pairs of groups in terms of observed variables that researchers may consider potential confounders. Looking at the size of the differences in Table 7, there is only 1 among the 52 categories or variables analysed[109] in which there was an imbalance greater than the 0.1 threshold recommended by Stuart et al. (2013) and Austin (2009). For this imbalanced variable ("age" in D_1), we also examined its unstandardized mean difference following Rosenbaum (2010). Specifically, the average difference of age among the groups of D_1 is of only one year or, equivalently, 0.13 standard deviations. Given the small standardized difference and the practical impact of one year on the employment probability, we argue that the confounding effect of this slight age difference is negligible for the conclusions of our study.

The next step is to subject our models to robustness tests to see whether the conclusions change when we vary different parameters. We are going to present only the main results of these checks, with detailed tables available in Appendix E for further reference. Let us see the results for each estimand. For the $LATE$, we altered two modelling options: the bandwidth selector and the polynomial degree. We tried two alternative bandwidth selectors, both of them pursuing the minimization of the mean squared error: the same procedure but with different bandwidths at each side of the cutoff and the common bandwidth chose for the main analysis but with a small perturbation factor that makes it smaller. The polynomial form was either linear or quadratic. Once we apply these variations, our estimates indicate that the results are very robust to different polynomial degrees and almost equal in function to the bandwidth. There are some minor differences that might be generated by the statistical power of the distinct samples.

[109] Each treatment comparison analysed 26 categories/variables.

Table 7 – Covariate balance analysis for each treatment comparison

	D_1	D_2
Age	-1.071 (0.1336)	-0.499 (-0.0703)
Sex (Woman)	-0.061	-0.019
Foreign (Yes)	-0.037	0.017
Disability (True)	0.013	-0.005
Level of educational attainment		
Primary (*Elementare*)	-0.001	0.001
Lower secondary (*Media*)	0.028	0.014
Upper secondary of 3 years (*Qualifica*)	-0.010	0.007
Upper secondary of 4 years (*Diploma*)	-0.051	-0.037
University or equivalent (*Laurea*)	0.046	0.004
Last occupation		
Intellectual (2)	-0.002	0.000
Technical (3)	0.010	-0.005
White collar low-skilled (4)	-0.005	-0.002
White collar high-skilled (5)	-0.034	0.002
Blue collar high-skilled (6)	-0.008	0.035
Blue collar intermediate-skilled (7)	0.001	-0.003
Blue collar low-skilled (8)	0.000	-0.064
Other occupations or unemployed (1, 9)	0.025	0.044
Sector of last employment		
Made in Italy	-0.050	-0.011
Other industries	0.034	-0.010
Construction	0.004	-0.016
Retail and leisure	-0.014	0.011
Wholesale and logistics	0.042	0.029
Advanced third sector	-0.017	-0.011
Other services	0.006	-0.002
Personal services	-0.008	0.007
Agriculture or unobserved	0.002	0.042

Note: The cell shows the estimated difference (of means or proportions) between both groups for each comparison. The estimator used a polynomial degree one and triangular kernel. The quantitative variable, "age", also includes the standardized mean difference in parentheses. The categorizations of level of education, occupations, and sectors follow the statistical classifications of Italy and Veneto.

The other estimators present less established procedures to test the robustness, because they are relatively newer[110]. For the $LQTE(\tau)$, we have also altered the bandwidth according to the proposal of Qu & Yoon (2019), i.e. selecting the second instead of the first most conservative bandwidth. In the two treatment contrasts, the findings are qualitatively equivalent with slight differences in the effect size. In any case, the conclusions we extract from these analyses are the same. For the $LATE_j$, we have varied the bandwidth selection procedure by generating one binary variable for each category and choosing the bandwidth for each dummy in an independent way. In this robustness check, the direction of the effects is the same and the coefficients barely change with respect to the ones presented in the main analysis. The point estimates for these alternative estimates are included in the confidence intervals offered in the previous section, so the conclusions are the same.

Lastly, due to the possibility of heterogeneous effects, we have conducted subgroup analyses. We have focused on subpopulations or subgroups defined by one of these four variables: sex, age, level of educational attainment, and implementation during COVID-19. While some of these factors present a clear categorization (e.g. sex or COVID-19), others involve less obvious categorizations. For the age variable, we have separated the individuals older and younger than 50 years olds due to the simultaneous existence of a policy in Italy that subsidises social security contributions of workers above that age with long-term unemployment (Pacifico et al., 2018). For the education variable, we have first split those with compulsory and non-compulsory levels of education. Then, we have separated those with upper secondary education and those with tertiary education due to their different occupational specificity.

[110] For instance, there has not been developed a procedure to decide which functional form to choose if the results substantially differ in the style of Pei et al. (2022) but for the quantile setup.

Table 8 – Subgroup analyses for D_1

	Y_1	Y_2	Y_3	Y_4	Y_{OEC}	Y_{FTCm12}
Sex = Woman	-7.898	-3.234	8.415	5.077	-0.025	-0.004
	(4.700)	(5.615)	(7.372)	(6.625)		
Sex = Man	0.100	9.199	7.736	3.546	0.043***	0.003
	(5.014)	(5.452)	(6.181)	(6.400)		
Age ≥ 50	6.889	15.278	13.715	-0.600	0.007	0.006
	(7.128)	(7.969)	(8.656)	(9.914)		
Age < 50	-4.687	1.886	6.614	5.660	0.016	-0.002
	(4.267)	(5.516)	(6.315)	(5.900)		
Educ. = 1 or 2	-3.349	3.769	0.939	-2.571	0.047**	-0.002
	(6.405)	(5.279)	(6.430)	(7.028)		
Educ. = 3	3.320	12.579*	24.879***	19.108**	0.005	0.001
	(6.088)	(7.407)	(9.205)	(8.733)		
Educ. = 6, 7, or 8	-10.195	-9.425	-0.790	-2.967	-0.012	-0.001
	(8.423)	(10.601)	(10.929)	(9.759)		
Date < COVID-19	-4.643	0.971	.	.	.	.
	(5.897)	(7.642)	.	.		
Date ≥ COVID-19	-7.704	1.915	2.072	-5.401	-0.056	0.003
	(11.022)	(12.919)	(12.544)	(11.186)		

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The row indicates the subpopulation. In the columns of quantitative outcomes, each cell includes the point estimate and the standard error of a model with a linear term for the scoring variable and triangular kernel. In the columns of qualitative outcomes, the cell only offers the point estimate. Level of education measured through ISCED. The date that sets the starting of the COVID-19 subpopulation is 01/01/2020. The subpopulation not affected by COVID-19 is made up by those that started AXL before 01/01/2019. Results on the third and fourth semester are not shown since at these periods COVID-19 was already active.

Table 8 offers the results of the subgroup analyses for treatment comparison D_1. The general conclusions extracted from the pooled analysis also hold in most cases here, since the estimates are mostly not statistically significant. There are three exceptions that merit attention. Firstly, the subgroup of jobseekers with upper secondary education presents an increase of around 20-25 worked days for the third and four semesters when assigned to group B instead of group A. If we examine this result more closely, we observe that the impact of the greater endowment on the number of hours of assistance (Y_{HS}) or adult training (Y_{HT}) follows the same pattern previously detected: in average, one extra hour of assistance and same hours of training. Thereby, the larger effect might not be explained by a longer treatment. A possible reason for this effect might be that providers concentrate increments of efforts in those skill groups that know better and already have any vocational education.

Secondly, looking at the employment quality[111], the subgroup of workers with compulsory education presents a 5 p.p. rise in the probability of signing an open-ended contract as their longest employment episode. This contrasts with the null effects found at the number of worked days. A possible explanation is that this contracts did not last as expected due to disciplinary dismissals or low-quality companies with economic problems. Thirdly, there is also a positive effect of 4 p.p. on Y_{OEC} for the subpopulation of men. This might be explained by the larger prevalence of people with level of education 1 or 2 in men (51 % for the whole sample of D_1) than in women (26.8 % for the same sample).

The results for the second comparison (D_2) are even more consistent with the findings of the pooled sample. As we see in Table 9, almost all coefficients are not statistically significant at the usual $\alpha = 0.05$ level. We do not detect differences in the number of worked days when going from B to group C in the subgroup analyses of sex, age, level of education, or COVID-19. Turning to the employment quality variables, we find only one statistically significant coefficient. We estimate that changing from group B to group C has an effect of 3.7 p.p. less of probability of signing an OEC as the longest contract for jobseekers younger than 50. If we review the results for the additional types of contracts presented in Appendix E, we can appreciate that such decrease was compensated by an increase in the probability of signing a short temporary contract during the longest employment.

[111] The findings for the other two type of contracts (FTC612 and FTC05) are in Appendix D.

Table 9 – Subgroup analyses for D_2

	Y_1	Y_2	Y_3	Y_4	Y_{OEC}	Y_{FTCm12}
Sex = Woman	7.280	8.356	2.805	-0.913	-0.031	0.001
	(5.638)	(6.649)	(6.630)	(6.457)		
Sex = Man	2.122	6.941	0.033	3.655	-0.007	0.003
	(4.776)	(6.122)	(5.468)	(6.162)		
Age ≥ 50	-5.238	-0.664	-1.338	7.341	-0.028	0.006
	(5.980)	(6.165)	(7.493)	(9.486)		
Age < 50	8.423*	8.553	3.902	2.106	-0.037**	0.002
	(5.084)	(5.566)	(5.552)	(5.407)		
Educ. = 1 or 2	11.735*	9.432	2.941	-2.902	-0.023	-0.003
	(7.544)	(7.130)	(6.523)	(6.863)		
Educ. = 3	3.372	3.969	0.070	10.085	-0.030	0.006
	(5.160)	(6.308)	(6.579)	(7.339)		
Educ. = 6, 7, or 8	-0.131	3.027	-6.416	3.541	-0.031	-0.005
	(8.117)	(10.499)	(12.650)	(12.515)		
Date < COVID-19	8.108	8.357	.	.	.	.
	(6.323)	(6.746)	.	.		
Date ≥ COVID-19	7.532	0.691	-7.754	-12.612	-0.029	0.002
	(10.832)	(11.099)	(10.900)	(13.423)		

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The row indicates the subpopulation. In the columns of quantitative outcomes, each cell includes the point estimate and the standard error of a model with a linear term for the scoring variable and triangular kernel. In the columns of qualitative outcomes, the cell only offers the point estimate. Level of education measured through ISCED. The date that sets the starting of the COVID-19 subpopulation is 01/01/2020. The subpopulation not affected by COVID-19 is made up by those that started AXL before 01/01/2019. Results on the third and fourth semester are not shown since at these periods COVID-19 was already active.

5. Conclusion

In this chapter, we have examined the employment effects of two treatment endowments of an ALMP developed in the region of Veneto (Italy) from 2017 until 2022. The program assigned the jobseekers on the basis of an employability score to three treatment groups sorted with different endowments of job search assistance and intermediation: group A (7 hours of assistance plus 1,500 € for intermediation), group B (13 hours of assistance plus 2,000 € for intermediation), or group C (27 hours of assistance plus 3,000 € for intermediation).

Using a regression discontinuity analysis, we have estimated the local effect of participating in a certain group (the active treatment) compared to the group immediately preceding it in intensity (the control treatment). Regarding employment duration, the effects on the average are null for both treatment comparisons (B vs. A and C vs. B) one year after participation started. Some effects appear during the first semesters, but later they disappear. The effects on the median of the distribution of worked days have a similar pattern. There are some impacts of around one month on certain semesters, but at the end of the second year there is not evidence on these differences. The impacts on the first and third quartiles of the distribution are always absent. Regarding employment quality, we detect a statistically significant positive effect on the probability of signing a short temporary contract as the longest contract in two years for the second treatment comparison (C vs. B). For the same comparison, we also find a decrease in the probability of having a permanent contract as the longest contract in two years. The analysis on the mechanism of treatment uptake revealed that there were minor differences in hours of assistance between groups. If vouchers with distinct endowments were finally transformed in the same hours of job search assistance, the unique channel to boost a differential effect would be higher efforts in intermediation. Additionally, the plausibility and robustness checks overall supported the analytical strategies. With the heterogeneity analyses, we may highlight that the first treatment comparison did have a positive effect on the number of worked days for the jobseekers with upper secondary level of education. It seems that such increases in the efforts of job search intermediation were concentrated on this type of workers, although we do not detect such behaviour for the second treatment comparison (C vs. B).

In relation to the literature, our study has tried to make progress on two lines. On one side, we have been the first to offer an estimate of the effect of different intensities of job search services inside the same program in Europe. Previous research had compared the effect of different formats of assistance or intermediation by comparing individual versus group meetings (Maibom et al., 2017; Bollens & Cockx, 2017). Consistent with the literature for USA, we have also found null effects of the intensity on the employment probability (Liu et al., 2014). On the other side, we have credibly estimated effects under a voucher implementation. In line with previous research for Italy, we have also reported a low compliance with the endowments of vouchers (Zanella & Salomone, 2024; Linfante et al., 2019). All our treatment groups allowed a partial expenditure of the voucher (what we called open-dose), so we could not contrast their results with a closed-dose voucher. Still, our null results on employment effects are also similar to the ones found in Brenninkmeijer & Blonk (2012) for an open-dose voucher of ALMP in Netherlands.

There are some insights obtained from our work that might be applied in the next policy designs. Concerning assistance, we show that this open-dose version of vouchers does not ensure that jobseekers with higher unemployment risk receive a more intense treatment. Future versions of the program might consider closed-dose vouchers in which the jobseeker may select the provider, but not the number of hours of assistance. It would also be interesting to know the reasons of the participants about why they did not continue the treatment. Regarding intermediation, it is possible that the differences among the payments linked to each treatment group were not sufficient to foster the efforts required to trigger an effect. Since effort is unobserved for the PES, an alternative explanation might be that providers prefer to devote their working hours to services like assistance that are not performance-based. In the future, the Veneto PES might ask providers for reporting information on the actions related to job search intervention.

The research has some limitations that should be considered. In terms of identification, regression discontinuity design offers credible identification assumptions at the cost of an estimand that only deals with people at the cutoff of the running variable. The main downside is that we need to assume homogeneous effects in function of the score if we want to extrapolate the

estimates to individuals far from the threshold[112]. Moreover, with these lenient assumptions we have identified the effects of two treatment *assignments*. The freedom to choose how much of the vouchers to spend resulted in endowments of different amounts ultimately leading to similar doses of treatment received. On the one hand, this lack of compliance with the available doses of assistance might lead us to interpret the estimates as the impacts of different payment schemes of job search intermediation. On the other hand, as was done in other studies on vouchers, we could take the results as estimates of intention-to-treat effects. This is our favourite option, since the partial expenditure of vouchers may be thought as a negative effect of this program design. In terms of data, we must remind that we worked with the available datasets provided by the PES of Veneto. Our outcome variable on the number of worked days only collects information on salaried employment, so there is not self-employment. This might induce us to think that we are actually collecting a lower bound of the treatment effect, since the program might also affect the probability of self-employment.

There are at least two open paths for further research. On the identification problem, researchers might carry out impact evaluations using a control group made up by individuals without any treatment. In this case, the identification assumptions are usually stronger, but the panel structure of the data allows the possibility to estimate models collecting unobservable but constant confounders. On the data problem, future research might explore the acquisition of social security datasets or more detailed data on the treatment. This fine-grained definition of treatments has been promoted by causal inference researchers from other disciplines like Hernán (2016). For instance, there are opportunities to improve the description of actions labelled as intermediation to explore which type of these actions did not generate the expected results.

The bottom line is that, despite the limitations, our evaluation shows that the differences between endowments of job search assistance and intermediation did not generate effects that lasted until the second year. In addition, we have evidenced that larger assignments did not substantially change the number of treatment doses. Therefore, our findings might help to rethink some aspects of the voucher and the ALMP quasi-market.

---

[112] We tried the strategy of Angrist & Rokkanen (2015), but the results did not support the extrapolation of findings beyond the threshold conditional on our available covariates.

Appendix

Appendix A. Cream-skimming analysis

Provider	D_1	D_2
GI GROUP	-0.013	0.008
	(-0.029, 0.014)	(-0.005, 0.029)
UMANA	0.006	-0.013
	(-0.014, 0.038)	(-0.043, 0.004)
ADECCO ITALIA	0.006	0.011
	(-0.016, 0.02)	(-0.009, 0.024)
Manpower	0.001	-0.006
	(-0.013, 0.024)	(-0.026, 0.007)
ENAIP VENETO	-0.008	-0.016
	(-0.026, 0.002)	(-0.03, 0.006)
RANDSTAD ITALIA	0.007	-0.005
	(-0.005, 0.028)	(-0.019, 0.003)
Synergie Italia	-0.017***	-0.01
	(-0.033, -0.007)	(-0.02, 0.005)
Staff	-0.001	-0.009
	(-0.009, 0.013)	(-0.025, 0.003)
Fondazione ENAC Veneto	0	0.003
	(-0.01, 0.008)	(-0.007, 0.009)
Infolingue	0.006	-0.004
	(-0.004, 0.012)	(-0.013, 0.002)
APINDUSTRIA SERVIZI	0.001	-0.006
	(-0.014, 0.019)	(-0.013, 0.004)
Eurointerim	0.004	
	(-0.007, 0.01)	
Eduforma		0.011***
		(0.004, 0.022)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate of the $LATE_j$ and the 95 % confidence interval in parentheses. To avoid small sample sizes, we only included the set of providers that accounted for 75% of the sample (having previously sorted the providers from more to less users of AXL).

Appendix B. On the variables

B.1. Outcome variables

Worked days as salaried employee at semester *t* (quantitative). This variable gathers the number of worked days as salaried employment, not as self-employment, in a certain semester. Specifically, we observe it as semesters +1, +2, +3, and +4 since the beginning of the treatment. Notice that it only registers the days actually worked or, in other words, the working days under active labour relationships.

Type of labour contract of the longest employment (qualitative, ordered). This variable collects the type of labour contract associated to the employment episode with the longest expected duration during the time interval initiated from the beginning of the treatment to 24 months later. Open-ended contracts have an expected duration of ∞ days. This variable has five possible values: open-ended contract, fixed-term contract with expected duration of more than 12 months, fixed-term contract with expected duration of [6, 12] months, fixed-term contract with expected duration of less than 6 months, no employment. Then, it is a variable that gathers the expected duration of the contract regardless of the specific Italian name of such contract. For instance, this means that the category "open-ended contract" not only includes "*contratto tempo indeterminato*", but also the open-ended version of the "*contratto domestico*".

B.2. Confounding variables

Scoring variable (quantitative). The score output by the statistical profiling model used by the Veneto PES to allocate people into treatment groups. It is a function of the following variables: sex, age, foreign nationality, level of studies, having a disability, termination reason of the last job, major group of occupation of the last job, and sector of the company of the last job.

B.3. Additional variables

Age (quantitative). Number of years old at the date the individual got the voucher of AXL.

Sex (qualitative, unordered). Sex of the participant, either man or woman.

Foreign (qualitative, unordered). It collects if the individual has the Italian citizenship or not.

Disability (qualitative, unordered). It gathers if the participant has any type of disability.

Level of educational attainment (qualitative, unordered). It includes the Italian name of the credential assigned to the highest level of educational attainment of the participant. We can then convert it to the ISCED-A classification of levels of education.

Last occupation of last job (qualitative, unordered). It registers the occupation that the individual had in its last labour contract before participating in AXL. Originally, it was written using the Italian classification of occupations called CP2011. We may then convert it to ISCO-08.

Industry of last job (qualitative, unordered). This variable collects the industry of the last job using the ATECO 2007, the Italian classification of economic activities. To simplify it, we have converted it to the second level of the classification of economic activities used in Veneto Lavoro (the Veneto PES), which may be found in https://www.venetolavoro.it/glossario-e-strumenti.

C. Sample statistics

Table C1 – Summary statistics on pretreatment variables

	Group A	*Group B*	*Group C*	*All*
Age	40	43.7	48.8	44
Sex (Woman)	43.1	50.9	59.7	50.9
Foreign (Yes)	17.7	20.6	29.1	22.3
Disability (Yes)	3	5.3	10.1	5.9
Level of educational attainment				
Primary	2.5	3.4	7.2	4.3
Lower secondary	34.0	41.2	50.9	41.7
Upper secondary general	9.8	8.3	7.0	8.4
Upper secondary vocational	35.2	33.2	27	32
University or equivalent (arts included)	18.5	13.9	7.8	13.6
Last occupation				
High skilled white collar (ISCO 1, 2 and 3)	11.9	8.9	5.4	8.9
Low skilled white collar (ISCO 4 and 5)	31.5	27.2	20	26.5
High skilled blue collar (ISCO 6 and 7)	12.5	12.2	11.8	12.2
Low skilled blue collar (ISCO 8 and 9)	32.4	29.8	25.1	29.2
Unemployed	11.6	21.9	37.8	23.2
Number of observations	11,492	10,782	10,171	32,445

Note: Each cell shows the column proportion (qualitative variables) or the mean (quantitative variables). Since this table might be useful to transport our results to other contexts (for external validity), we offer the level of studies according to ISCED-A-11 and occupations according to ISCO-08.

Figure D1 – RD plots with local means for employment duration variables (semesters 3 and 4)

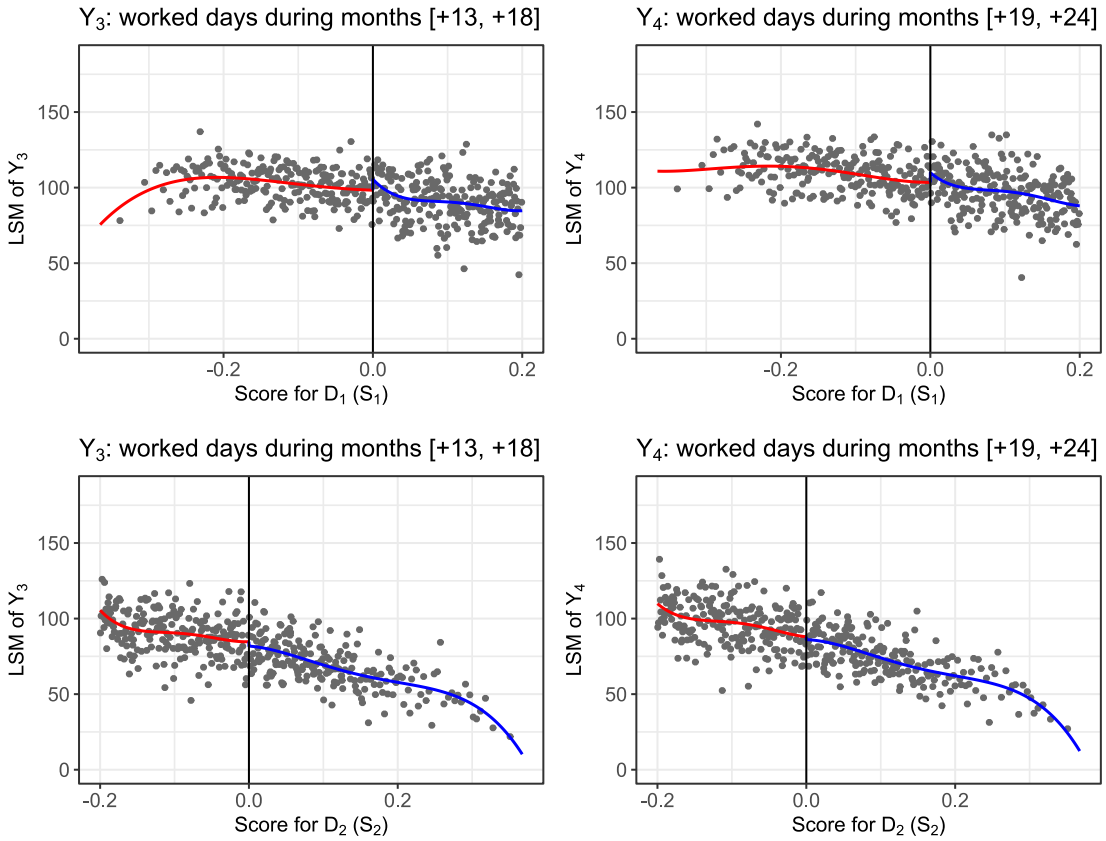


Figure D2 – RD plots with local medians for employment duration variables (semesters 3 and 4)

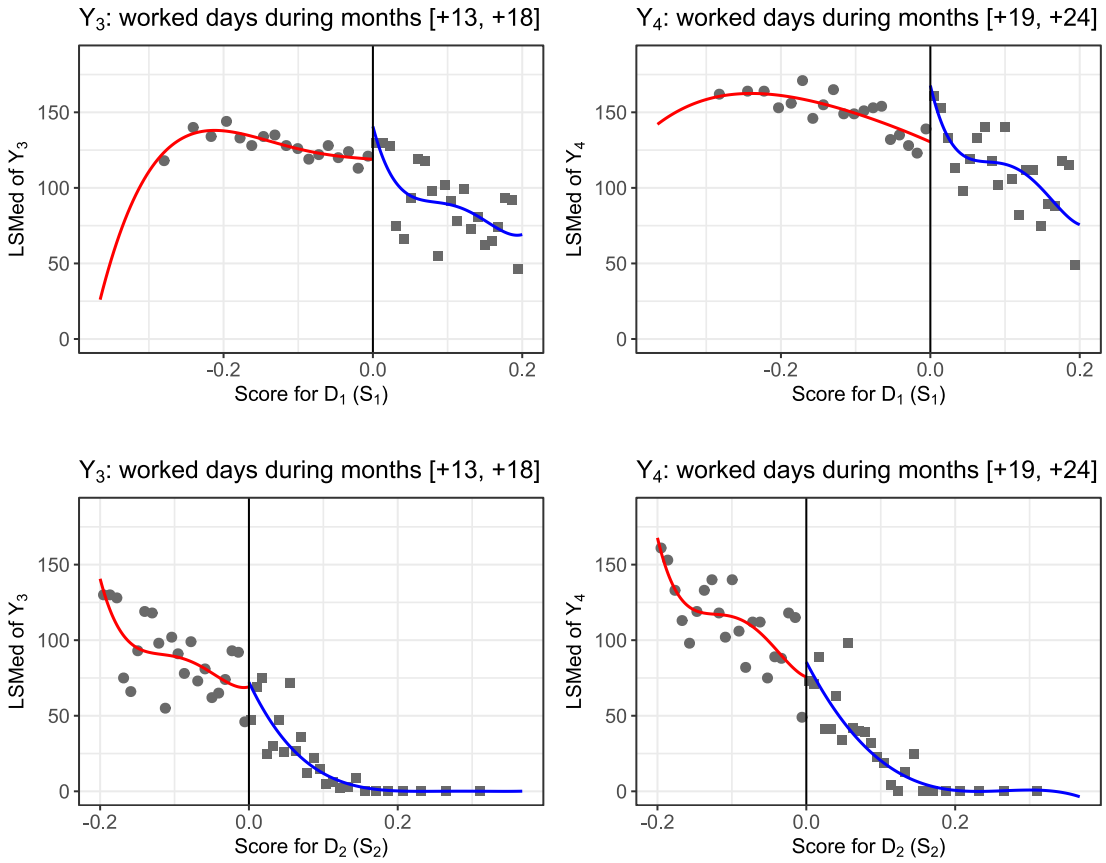
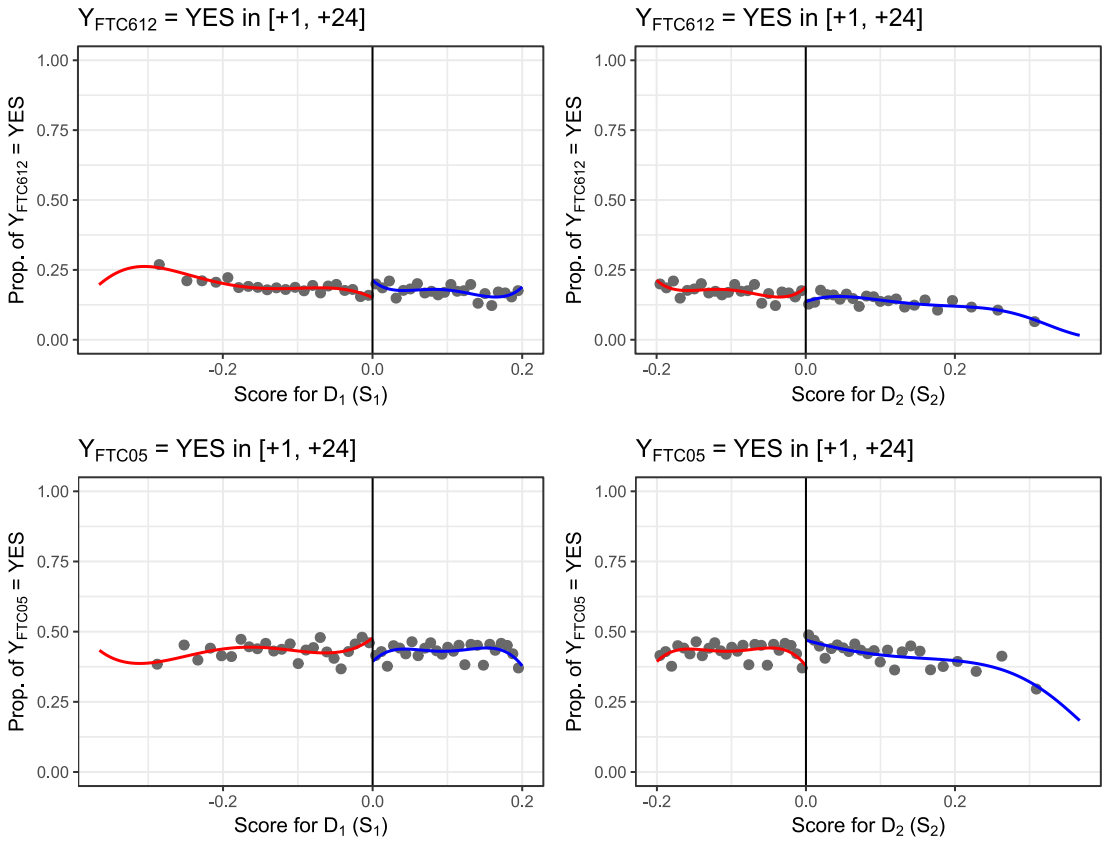


Figure D3 – RD plots with local proportions for employment quality variables (FTC612 and FTC05)



E. Additional tables

Robustness checks for $LATE$

Table E1. Estimates of $LATE$ of D_1 on Y_1 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**-1.745**	-1.665	-1.750	-1.416
	(3.715)	(4.847)	(4.051)	(5.187)
$\widehat{AMSE}$	15.159	24.426		
Bandwidth	0.063	0.08	0.053	0.07
N_h	6564	8304	5548	7253
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E2. Estimates of $LATE$ of D_1 on Y_2 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**4.837**	4.336	4.579	5.235
	(4.174)	(4.864)	(4.560)	(5.127)
$\widehat{AMSE}$	17.633	24.021		
Bandwidth	0.062	0.097	0.052	0.087
N_h	6400	10128	5412	9167
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E3. Estimates of $LATE$ of D_1 on Y_3 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**8.075**	9.995*	7.214	9.721
	(5.115)	(5.588)	(5.776)	(5.983)
$\widehat{AMSE}$	28.227	32.183		
Bandwidth	0.045	0.081	0.035	0.071
N_h	4719	8393	3749	7324
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E4. Estimates of $LATE$ of D_1 on Y_4 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**4.587**	4.224	2.718	3.479
	(4.738)	(5.467)	(5.272)	(5.830)
$\widehat{AMSE}$	22.252	30.091		
Bandwidth	0.053	0.085	0.043	0.075
N_h	5507	8784	4430	7722
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E5. Estimates of $LATE$ of D_2 on Y_1 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**6.835**	8.415*	4.502	7.794
	(4.387)	(5.090)	(5.189)	(5.642)
$\widehat{AMSE}$	21.476	26.500		
Bandwidth	0.037	0.061	0.027	0.051
N_h	4079	6614	3032	5566
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E6. Estimates of $LATE$ of D_2 on Y_2 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**10.378****	14.512**	11.907	13.347
	(4.929)	(6.455)	(5.665)	(7.157)
$\widehat{AMSE}$	27.995	44.693		
Bandwidth	0.043	0.056	0.033	0.046
N_h	4741	6151	3635	5077
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E7. Estimates of $LATE$ of D_2 on Y_3 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**4.201**	11.135*	7.314	11.458
	(4.806)	(7.114)	(5.472)	(8.038)
$\widehat{AMSE}$	25.522	54.919		
Bandwidth	0.048	0.05	0.038	0.04
N_h	5246	5499	4116	4452
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E8. Estimates of $LATE$ of D_2 on Y_4 with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**7.512**	15.607**	11.531	15.707
	(5.341)	(7.687)	(6.189)	(8.788)
$\widehat{AMSE}$	32.247	66.321		
Bandwidth	0.043	0.047	0.033	0.037
N_h	4640	5157	3573	4022
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E9. Estimates of $LATE$ of D_1 on Y_1 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**-2.424**	-2.036
	(3.604)	(4.757)
$\widehat{AMSE}$	8.447 5.665	10.274 13.151
Bandwidth	0.062 0.075	0.101 0.07
N_h	7082	8800
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E10. Estimates of $LATE$ of D_1 on Y_2 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**4.775**	4.407
	(4.254)	(4.739)
$\widehat{AMSE}$	8.641 11.466	10.951 12.422
Bandwidth	0.071 0.051	0.114 0.093
N_h	6338	10630
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E11. Estimates of $LATE$ of D_1 on Y_3 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**7.980**	10.370*
	(4.496)	(5.180)
$\widehat{AMSE}$	5.569 15.408	9.314 18.179
Bandwidth	0.106 0.041	0.134 0.074
N_h	7531	10448
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E12. Estimates of $LATE$ of D_1 on Y_4 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**5.515**	3.797
	(4.785)	(5.662)
$\widehat{AMSE}$	9.642 15.088	17.132 17.261
Bandwidth	0.069 0.042	0.081 0.077
N_h	5787	8282
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E13. Estimates of $LATE$ of D_2 on Y_1 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**6.300***	9.931**
	(3.929)	(4.926)
$\widehat{AMSE}$	14.289 2.783	18.794 6.578
Bandwidth	0.033 0.117	0.054 0.1
N_h	7881	8228
Kernel	Triangular	Triangular
Polynomial degree	1	2

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E14. Estimates of $LATE$ of D_2 on Y_2 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**6.003**	14.820**
	(4.231)	(6.234)
$\widehat{AMSE}$	14.282 5.236	29.176 12.226
Bandwidth	0.045 0.1	0.051 0.086
N_h	7780	7352
Kernel	Triangular	Triangular
Polynomial degree	1	2

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E15. Estimates of $LATE$ of D_2 on Y_3 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**4.229**	13.267**
	(4.542)	(6.738)
$\widehat{AMSE}$	16.882 5.64	39.37 9.799
Bandwidth	0.041 0.093	0.044 0.11
N_h	7175	8109
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E16. Estimates of $LATE$ of D_2 on Y_4 with one MSE-optimal bandwidth selector for each side

	(1)	(2)
Treatment	**7.485**	16.483**
	(4.982)	(7.083)
$\widehat{AMSE}$	21.464 6.265	44.966 10.431
Bandwidth	0.038 0.088	0.042 0.112
N_h	6753	8068
Kernel	Triangular	Triangular
Polynomial degree	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table E17. Estimates of $LATE$ of D_1 on Y_{HS} with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**1.252*** **	1.141***	1.213***	1.162***
	(0.211)	(0.240)	(0.239)	(0.258)
$\widehat{AMSE}$	0.050	0.060		
Bandwidth	0.045	0.076	0.035	0.066
N_h	4685	7870	3741	6803
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E18. Estimates of $LATE$ of D_1 on Y_{HT} with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**-0.216**	-1.078	-0.330	-1.489
	(1.331)	(2.176)	(1.413)	(2.332)
$\widehat{AMSE}$	1.766	4.794		
Bandwidth	0.088	0.067	0.078	0.057
N_h	9233	6964	8173	5953
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E19. Estimates of $LATE$ of D_2 on Y_{HS} with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**1.110*** **	1.162***	1.098***	1.218**
	(0.256)	(0.380)	(0.275)	(0.412)
$\widehat{AMSE}$	0.065	0.146		
Bandwidth	0.074	0.074	0.064	0.064
N_h	7973	8023	6918	6948
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Table E20. Estimates of $LATE$ of D_2 on Y_{HT} with one common MSE-optimal bandwidth and estimates with a slight change of this bandwidth.

	(1)	(2)	(3)	(4)
Treatment	**2.209**	1.974	2.550*	2.154
	(1.724)	(1.913)	(1.906)	(2.026)
$\widehat{AMSE}$	2.912	3.670		
Bandwidth	0.055	0.097	0.045	0.087
N_h	6011	10,432	4982	9344
Kernel	Triangular	Triangular	Triangular	Triangular
Polynomial degree	1	2	1	2

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Columns 1-2 use a MSE-optimal bandwidth.

Robustness checks for $LQTE(\tau)$

Table E21. Estimates of $LQTE(\tau)$ with the second lowest bandwidth

	Y_1	Y_2	Y_3	Y_4
D_1				
$\tau = 0.25$	0	0	0	0
$\tau = 0.5$	-3.158	4.357	16.640	32.409
$\tau = 0.75$	4.059	0	0.040	1.092
Treatment significance test	.	.	***	.
Treatment homogeneity test	.	.	***	.
Bandwidth	0.088	0.09	0.071	0.048
D_2				
$\tau = 0.25$	0	0	0	0
$\tau = 0.5$	3.875	27.777	-6.87	-3.327
$\tau = 0.75$	-5.783	1.977	0	0
Treatment significance test	***	***	.	.
Treatment homogeneity test	**	***	.	.
Bandwidth	0.086	0.057	0.119	0.116
Bandwidth estimator		2^{nd} lowest		
Kernel		Epanechnikov		

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Robustness checks for $LATE_j$

Table E22 – Estimates of employment quality effects for each treatment and contract using one different bandwidth for each category

	Y_{OEC}	Y_{FTCm12}	Y_{FTC612}	Y_{FTC05}
D_1	0.021*	-0.001	0.027	-0.029*
	(-0.003, 0.061)	(-0.01, 0.005)	(-0.007, 0.056)	(-0.079, 0.004)
D_2	-0.029*	0.001	-0.006	0.045**
	(-0.073, 0)	(-0.003, 0.008)	(-0.036, 0.012)	(0.012, 0.095)

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate of the $LATE_j$ and the 95 % confidence interval in parentheses.

Table E23 - Subgroup analyses for D_1, rest of categories of the employment quality outcome

	Y_{FTC612}	Y_{FTC05}
Sex = Woman	0.042*	-0.018
Sex = Man	0.002	-0.007
Age ≥ 50	0.046***	-0.007
Age < 50	0.014	-0.026
Educ. = 1 or 2	0.009	-0.035
Educ. = 3	0.029	-0.019
Educ. = 6, 7, or 8	0.003	0.052
Date < Covid	.	.
Date ≥ Covid	0.022	0.038

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate of the $LATE_j$.

Table E24 - Subgroup analyses for D_2, rest of categories of the employment quality outcome

	Y_{FTC612}	Y_{FTC05}
Sex = Woman	0.005	0.029
Sex = Man	-0.025**	0.041**
Age ≥ 50	-0.010	0.076**
Age < 50	0.004	0.036**
Educ. = 1 or 2	-0.015	0.042*
Educ. = 3	-0.019	0.070***
Educ. = 6, 7, or 8	0.069*	-0.051
Date < Covid	.	.
Date ≥ Covid	-0.022	0.063

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Each cell shows the point estimate of the $LATE_j$.

Chapter V.

Conclusions

This doctoral dissertation has tried to provide evidence to improve the design of ALMPs relying on different methods and looking into different types of ALMPs. On one side, we have evaluated the effectiveness of some features of the *tools* used in ALMPs. Chapter II assessed the importance of the timing, the intensity, and the specificity of adult training. There is not clear evidence of a cohort effect, but there is a remarkable non-linear effect of the intensity of work-based training (internships) on the employment probability. However, the intensity or the occupational specificity of the whole treatment seem to be less important design factors. Chapter IV estimated the effect of different endowments of job search services, since this time the ALMPs were offered through vouchers and performance-based payments. Comparing different versions of the treatment, we detected that the assignment had no effect neither on the worked days nor on the quality of employment after two years. A key point was that the vouchers on assistance were not exhausted in average, resulting in small differences in treatment uptakes regarding this tool. Moreover, the intermediation pays seemed to be insufficient to generate employment effects. On the other side, we have also evaluated in Chapter III the predictions of jobseeker profiling models to select the *target groups* of the ALMPs. To compare the results of the current rule-based profiling model used in Catalonia, we have developed an alternative statistical model. We have found that the rule-based procedure has a poor predictive performance both regarding to its discrimination ability and its calibration. On the contrary, our statistical model achieves a state-of-the-art performance with the current administrative data and attains an excellent calibration in its machine learning version. This may foster the creation of jobseeker groups closer to their actual situation in the future.

Each chapter contextualizes these results inside the academic literature, but we would also like to discuss how these findings might fit in the public policy sphere. On the effects dimension, combining all our results, the research supports a more *guided* design of ALMPs. In the continuum on how much decisions we charge on

the jobseeker or on the government, we argue that our evidence backs to tilt the balance more in favour of the government. Firstly, the importance of the internships inside adult training programs makes us think that governments should facilitate this transition between education and work. Despite there are independent programs of internships or employment subsidies, a policy mix that integrates both classroom-based and work-based training might be more useful due to different mechanisms. Again, informational asymmetries may prevent jobseekers to use these independent programs if they need to select a trustable employer both regarding the training and the employment probabilities inside her staff. The necessity to combine both versions of training is also supported by the concave form of the effect of internships on employment. This result supports the dual spirit of the new Vocational Training Act passed in Spain in 2022, but qualifies that it does not seem advisable to increase the proportion of internships beyond the 25 % (Camas Roda, 2022). Independent programs without classroom-based training may focus too much on a very specific education which may not be transportable to other companies. Thereby, our evaluation upholds that this combination between ALMPs was facilitated by the policy designers. Secondly, the voucher system has not provoked a greater use of resources by those more in need. Despite having the possibility, participants have not generally carried out a full redemption of vouchers. The reasons might be the complexity of understanding the quasi-market of ALMPs or the difficulties to appreciate an effect of these interventions, which would support treatment attrition. This attrition might be rational if the interventions are actually ineffective, which should not be immediately rejected according to recent research (Cheung et al, forthcoming). This invites PES to implement closed-dose interventions of job search services that had been evaluated in advance through small pilots.

On the predictions dimension, our findings show that the implementation of statistical profiling models to assist the allocation of treatments is a feasible task with the current administrative sources. Other approaches based on the evaluation of heterogeneous effects may be more fine-grained but require substantial changes in the treatment assignment mechanisms to get credible estimates. Our research has emphasized that allocation systems of individual into treatments are made up of two types of models: the profiling and the decision model. Statistical methods are useful for the first model, but organizational and philosophical questions that go beyond data need to be answered for the second model. A big challenge for

future policy work is to formalize in a detailed way this decision model and integrate it with the current or the new statistical profiling. Despite the remaining work, our current findings could inform the redesign of the allocation system of ALMPs currently in place in Italy under the GOL program (Bebber, 2024). This national program is similar to AXL, but combines statistical and rule-based profiling with a small degree of discretionary leeway granted to the caseworker. In light of our results, we suggest reviewing the rule-based procedure with other methods for creating target groups like impact evaluations of heterogeneous effects. Beyond the technological changes on the management of databases that would require the deployment of a statistical model, policymakers should also worry about the trust of caseworkers on the profiling model. If the profiling procedure wants to be used as a diagnostic tool in the form of an intervention, training sessions for caseworkers on the development of the models might be an interesting mean to foster the reception of the procedure.

This work also suffers from some limitations that should be considered. For all the dissertation, we have emphasized since the beginning the importance of doing evaluations in each country to control for the possible interactions between all labour market policies. However, these labour market policies may also change over time within a country. For example, we must consider this when reading the results of the FOAP evaluation. In the Spain of 2024, the employment protection legislation has suffered certain changes since 2022 (Rojo Torrecillas, 2022). Due to the increase of the restrictions to hire with temporary contracts, current analyses on the employment quality might focus on a comparison of full vs. part-time contracts. For each chapter of the thesis, the weaknesses of each research are detailed in each chapter. Here we emphasize just to some important ideas. Firstly, Chapter II departs from an unknown assignment mechanism of treatments. Despite using more credible techniques like difference-in-differences, we must acknowledge that there is a fundamental identification uncertainty with these techniques that could be mostly solved with an experiment. Secondly, Chapter III proposes a complex machine learning algorithm to assist the creation of target groups. If the PES wants to deploy such a model, we should consider that caseworkers might reject it due to the fear of trusting a black-box device. Thirdly, Chapter IV carries out a comparison of different treatments inside the same program. We must remind that the null effects are for contrasts of different levels of endowments. This does not mean that receiving the lowest endowment had no

effect compared to receiving nothing, since this is a hypothesis that we have not tested.

In each research chapter, we have also presented some lines of future research. Now we simply highlight three open streams that deal with description, prediction, and effect estimation. On description, we have explained in Chapter III the necessity of developing coding instructions to offer a detailed and standardized description of program designs and program implementations. Similar proposals have been made by Cronert (2023) when discussing the limitations of the current datasets of expenditures and participants in ALMPs gathered by the EU. On prediction, we have expressed the necessity to think alternative strategies to predict long-term unemployment in the young population. We do not have long labour market histories for these individuals, so we need to explore new collections of data to forecast their risks. On effect estimation, there is still a long way to go in comparing experimental estimates with observational estimates, as done in Lalonde (1986). Although this requires the collaboration of PES, this exercise would give much more credibility to the estimates carried out with observational data.

In conclusion, we have offered empirical evidence to improve the design of ALMPs by evaluating the employment effects of certain tools and by facilitating the selection of target groups through a statistical model that predicts long-term unemployment. Although ALMPs are complex interventions that usually mix different tools and present distinct goals, we believe that this more fine-grained characterization of the design allows a faster improvement of their effectiveness and their efficiency.

References

Abadie, A., & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics*, *10*(1), 465–503.

Accetturo, A., Albanese, G., De Paola, M., & Torrini, R. (2024). The North–South Gap: Economic Development and Public Intervention. *Italian Economic Journal*.

Alujas Ruiz, J. A. (2002). *Políticas activas de mercado de trabajo en España 1985-2000* [Doctoral thesis].

Andrén, T., & Andrén, D. (2006). Assessing the employment effects of vocational training using a one-factor model. *Applied Economics*, *38*(21), 2469–2486.

Andrén, T., & Gustafsson, B. (2004). Income effects from labor market training programs in Sweden during the 1980s and 1990s. *International Journal of Manpower*, *25*(8), 688–713.

Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331–1344.

Arel-Bundock, V., Greifer, N., & Heiss, A. (n.d.). How to interpret statistical models using marginaleffects in R and Python. *Journal of Statistical Software*.

Arni, P., & Schiprowski, A. (2015). *Die Rolle von Erwartungshaltungen in der Stellensuche und der RAV-Beratung—Teilprojekt 2: Pilotprojekt Jobchancen-Barometer* [IZA Research Report,].

Ashenfelter, O. (1975). The Effect of Manpower Training on Earnings: Preliminary Results. *Proceedings of the Twenty-Seventh Annual Winter Meetings of the Industrial Relations Research Association*.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 47–57.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*(1), 685–725.

Auciello, I., & Puente, S. (2023). Cambios recientes en el patrón de convergencia entre regiones. *Boletín Económico*, *2023/T1*, 17.

Auer, P., Efendioglu, U., & Leschke, J. (2005). *Active labour market policies around the world: Coping with the consequences of globalisation.* International Labour Office.

Austin, P. C. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Communications in Statistics-Simulation and Computation*, *38*(6), 1228–1234.

Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model: Relation to the variance and odds ratio of a continuous explanatory variable. *BMC Medical Research Methodology*, *12*, 1–8.

Austin, P. C., & Steyerberg, E. W. (2019). The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Statistics in Medicine*, *38*(21), 4051–4065.

Autor, D. H. (2009). Studies of labor market intermediation: Introduction. In *Studies of labor market intermediation* (pp. 1–26).

Autor, D. H., Dorn, D., & Hanson, G. H. (2016). The China shock: Learning from labor-market adjustment to large changes in trade. *Annual review of economics*, *8*(1), 205-240.

Ayaita, A., Grund, C., & Pütz, L. (2022). Job placement via private vs. Public employment agencies: Investigating selection effects and job match quality in Germany. *Schmalenbach Journal of Business Research*, *74*(2), 137–162.

Bach, R. L., Kern, C., Mautner, H., & Kreuter, F. (2023). The impact of modeling decisions in statistical profiling. *Data & Policy*, *5*, 32.

Baert, S., Neyt, B., Siedler, T., Tobback, I., & Verhaest, D. (2021). Student internships and employment opportunities after graduation: A field experiment. *Economics of Education Review*, *83*, 102141.

Baily, M. N., Tobin, J., Holt, C., Wachter, M., Juster, T., Friedman, B., Perry, G., Hall, R., Okun, A., Feldstein, M., Duesenberry, J., & Gordon, R. J. (1977). Macroeconomic Effects of Selective Public Employment and Wage Subsidies. *Brookings Papers on Economic Activity*, *1977*(2), 511.

Barnes, S.-A., Wright, S., Irving, P., & Deganis, I. (2015). *Identification of latest trends and current developments in methods to profile jobseekers in European public employment services: Final report.*

Barnow, B. S. (2009). Vouchers in US vocational training programs: An overview of what we have learned. *Journal of Labor Market Research (Zeitschrift Für Arbeitsmarktforschung*, *42*(1), 71–84.

Bebber, A. (2024). Digitalization and Labour Market: the Role of Profiling in Occupational Transitions. *Italian Labour Law e-Journal*, *17*(1), 105-129.

Bell, S. H., Orr, L. L., Blomquist, J. D., & Cain, G. G. (1995). *Program applicants as a comparison group in evaluating training programs: Theory and a test.* W. E. Upjohn Institute for Employment Research.

Bentolila, S., Cabrales, A., & Jansen, M. (2023). *Does Dual Vocational Education and Training Pay Off?*

Berg, G. J., Kunaschk, M., Lang, J., Stephan, G., & Uhlendorff, A. (2024). *Predicting Re-Employment: Machine Learning Versus Assessments by Unemployed Workers and by Their Caseworkers* (Issue 202403)). Institut für Arbeitsmarkt-und Berufsforschung (IAB), Nürnberg [Institute for Employment Research.

Bia, M., Flores, C. A., Flores-Lagunes, A., & Mattei, A. (2014). A Stata package for the application of semiparametric estimators of dose–response functions. *The Stata Journal*, *14*(3), 580–604.

Black, E., Raghavan, M., & Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–863.

Blanchard, O. J., Jaumotte, F., & Loungani, P. (2014). Labor market policies and IMF advice in advanced economies during the Great Recession. *IZA Journal of Labor Policy*, *3*(1), 2.

Blázquez, M., Herrarte, A., & Saez, F. (2019). Training and job search assistance programmes in Spain: The case of long-term unemployed. *Journal of Policy Modeling*, *41*(2), 316–335.

Blommaert, L., Muja, A., Gesthuizen, M., & Wolbers, M. H. J. (2020). The Vocational Specificity of Educational Systems and Youth Labour Market Integration: A Literature Review and Meta-Analysis. *European Sociological Review*, *36*(5), 720–740.

Boeri, T., & Ours, J. V. (2008). *The economics of imperfect labor markets*. Princeton University Press.

Bol, T., Ciocca Eller, C., Werfhorst, H. G., & DiPrete, T. A. (2019). School-to-work linkages, educational mismatches, and labor market outcomes. *American Sociological Review*, *84*(2), 275–307.

Bol, T., & Weeden, K. A. (2015). Occupational closure and wage inequality in Germany and the United Kingdom. *European Sociological Review*, *31*(3), 354–369.

Bollens, J., & Cockx, B. (2017). Effectiveness of a job vacancy referral scheme. *IZA Journal of Labor Policy*, *6*, 1–24.

Bonoli, G. (2007). Time matters: Postindustrialization, new social risks, and welfare state adaptation in advanced industrial democracies. *Comparative Political Studies*, *40*(5), 495–520.

Bonoli, G. (2010). The political economy of active labor-market policy. *Politics & Society*, *38*(4), 435–457.

Bonoli, G. (2011). Active labour market policy and social investment: A changing relationship. In N. Morel, B. Palier, & J. Palme (Eds.), *Towards a Social Investment Welfare State?: Ideas, Policies and Challenges* (pp. 181–204). Chapter, Bristol University Press.

Bonoli, G. (2013). *The origins of active social policy: Labour market and childcare policies in a comparative perspective*. Oxford University Press.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Brenninkmeijer, V., & Blonk, R. W. (2012). The effectiveness of the JOBS program among the long-term unemployed: A randomized experiment in the Netherlands. *Health Promotion International*, *27*(2), 220–229.

Brown, A. J., & Koettl, J. (2015). Active labor market programs-employment gain or fiscal drain? *IZA Journal of Labor Economics*, *4*, 1–36.

Burgess, S., Propper, C., Ratto, M., & Tominey, E. (2017). Incentives in the public sector: Evidence from a government agency. *The Economic Journal*, *127*(605), 117–141.

Busemeyer, M. R. (2017). Public Opinion and the Politics of Social Investment. In A. Hemerijck (Ed.), *The Uses of Social Investment* (Vol. 1). Oxford University Press.

Bustillo, R., Fernández-Macías, E., Antón, J. I., & Esteve, F. (2011). *Measuring more than money: The social economics of job quality*. Edward Elgar Publishing.

Caetano, C., & Callaway, B. (2023). *Difference-in-differences with time-varying covariates in the parallel trends assumption.*

Cahuc, P., Carcillo, S., & Zylberberg, A. (2014). *Labor economics*. MIT press.

Caliendo, M., Künn, S., & Mahlstedt, R. (2023). The Intended and Unintended Effects of Promoting Labor Market Mobility. *Review of Economics and Statistics*, 1–52.

Caliendo, M., Mahlstedt, R., & Mitnik, O. A. (2017). Unobservable, but unimportant? The relevance of usually unobserved variables for the evaluation of labor market policies. *Labour Economics*, *46*, 14–25.

Callaway, B. (2021). *Universal vs. Varying Base Period in Event Studies*. https://bcallaway11.github.io/posts/event-study-universal-v-varying-base-period

Callaway, B. (2023). Difference-in-Differences for Policy Evaluation. In K. F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Springer.

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. H. (2024). *Difference-in-differences with a continuous treatment* (Issue w32117)). National Bureau of Economic Research.

Callaway, B., & Karami, S. (2023). Treatment effects in interactive fixed effects models with a small number of time periods. *Journal of Econometrics*, *233*(1), 184–208.

Callaway, B., & Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230.

Calmfors, L. (1994). *Active labour market policy and unemployment: A framework for the analysis of crucial design features*. OECD Economic Studies.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, *17*(2), 372–404.

Calster, B., McLernon, D. J., Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). & Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. *BMC Medicine*, *17*(1), 230.

Camas Roda, F. (2022). La nueva Ley de Formación Profesional: en búsqueda de la simbiosis entre lo educativo y lo laboral. *Trabajo y derecho: nueva revista de actualidad y relaciones laborales*, (94), 2.

Card, D., Kluve, J., & Weber, A. (2018). What works? A meta analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, *16*(3), 894–931.

Card, D., & Sullivan, D. (1988). Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment. *Econometrica*, *56*(3), 497–530.

Casanova, J., Felgueroso, F., Pérez, J. I. G., & Jiménez-Martín, S. (2021). El perfilado estadístico como instrumento para la evaluación del impacto del programa Incorpora. In *Cuadernos económicos de ICE* (p. 102).

Cattaneo, M. D., Crump, R. K., Farrell, M. H., & Feng, Y. (2024). On binscatter. *American Economic Review*, *114*(5), 1488–1514.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Jansson, M., & Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, *115*(531), 1449–1455.

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, *14*, 821–851.

Cebrián López, I., Cueto Iglesias, B., & Mato Díaz, F. J. (2021). La evaluación de impacto de las políticas activas de empleo en España: El deseo y la realidad. *Cuadernos Económicos de ICE*, *102*.

CEDEFOP. (2014). *Terminology of European education and training policy*. https://www.cedefop.europa.eu/en/tools/vet-glossary

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheung, M., Egebark, J., Forslund, A., Laun, L., Rödin, M., & Vikström, J. (forthcoming). Does job search assistance reduce unemployment? Evidence on displacement effects and mechanisms. *Journal of Labor Economics*.

Choe, C., Flores-Lagunes, A., & Lee, S. J. (2015). Do dropouts with longer training exposure benefit from training programs? Korean evidence employing methods for continuous treatments. *Empirical Economics*, *48*, 849–881.

Choi, J. Y., & Lee, M. J. (2023). Score manipulation, density continuity and intent-to-treat effect for regression discontinuity. *Pacific Economic Review*, *28*(4), 552–569.

Clasen, J. (2013). Defining comparative social policy. In P. Kennett (Ed.), *A Handbook of Comparative Social Policy, Second Edition* (pp. 71–83). Edward Elgar Publishing.

Clasen, J., Clegg, D., & Goerne, A. (2016). Comparative social policy analysis and active labour market policy: Putting quality before quantity. *Journal of Social Policy*, *45*(1), 21–38.

Clasen, J., & Mascaro, C. (2022). Activation: A thematic and conceptual review. *Journal of European Social Policy*, *32*(4), 484–494.

Clegg, D., & Durazzi, N. (Eds.). (2023). *Handbook of labour market policy in advanced democracies*. Edward Elgar Publishing.

Clemente, J., García Castrillo, P., González-Álvarez, M., & Sanso, M. (2014). Una evaluación de la efectividad de la formación ocupacional para desempleados antes y después de la crisis económica: El caso de Aragón. *Hacienda Pública Española*, *208*(1), 77–106.

Colombo, E., Mezzanzanica, M., Montaletti, G., & Ronzoni, G. (2015). Counterfactual model for labour market evaluation policies. *ERSA Conference Papers*, *ersa15p1718)*.

Conde-Ruiz, J. I., García, M., Puch, L. A., & Ruiz, J. (2019). Calendar effects in daily aggregate employment creation and destruction in Spain. *SERIEs*, *10*, 25–63.

Corrales-Herrero, H., & Rodríguez-Prado, B. (2018). The role of non-formal lifelong learning at different points in the business cycle. *International Journal of Manpower*, *39*(2), 334–352.

Cremades Chueca, O. (2014). Los trabajos de colaboración social y sus límites. *IUSLabor. Revista d'anàlisi de Dret Del Treball*, *0*(2).

Crépon, B. (2018). Private providers of labor market services: A review of the evidence. In A. Bergström & L. Calmfors (Eds.), *Framtidens arbetsförmedling* (pp. 239–297).

Crépon, B., & Den Berg, G. J. (2016). Active labor market policies. *Annual Review of Economics*, *8*, 521–546.

Cronert, A. (2019). Varieties of Employment Subsidy Design: Theory and Evidence from Across Europe. *Journal of Social Policy*, *48*(4), 839–859.

Cronert, A. (2022). The multi-tool nature of active labour market policy and its implications for partisan politics in advanced democracies. *Social Policy and Society*, *21*(2), 210–226.

Cronert, A. (2023). Effects and explanations of active labour market policy: Theoretical and empirical challenges for cross-national research. In

Handbook of Labour Market Policy in Advanced Democracies (pp. 343–359). Edward Elgar Publishing.

Cuéllar-Martín, J., Martín-Román, Á. L., & Moral, A. (2019). An empirical analysis of natural and cyclical unemployment at the provincial level in Spain. *Applied Spatial Analysis and Policy*, *12*, 647–696.

Cueto, B., & López, F. (2019). The apprenticeship contract: An evaluation. *Hacienda Pública Española*, *231*, 15–39.

Cueto, B., & Suárez, P. (2015). El papel de las políticas activas: Una perspectiva desde las comunidades autónomas. *Ekonomiaz: Revista Vasca de Economía*, *87*, 282–309.

Cueto, B., Toharia, L., García Serrano, C., & Alujas, J. A. (2010). Los efectos de la formación ocupacional: ¿Importa la duración de las acciones? *Hacienda Pública Española*, *195*(4), 9–36.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

Dammert, A. C., Galdo, J., & Galdo, V. (2015). Integrating mobile phone technologies into labor-market intermediation: A multi-treatment experimental design. *IZA Journal of Labor & Development*, *4*(1), 11.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., & Sculley, D. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, *23*(226), 1–61.

Dasgupta, T., Pillai, N. S., & Rubin, D. B. (2015). Causal inference from 2K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *77*(4), 727–753.

Dauth, W., Findeisen, S., & Suedekum, J. (2017). Trade and manufacturing jobs in Germany. *American Economic Review*, *107*(5), 337-342.

Decision No 573/2014/EU of the European Parliament and of the Council of 15 May 2014 on Enhanced Cooperation between Public Employment Services (PES) (2021).

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, *94*(448), 1053–1062.

Den Berg, G. J., Hofmann, B., & Uhlendorff, A. (2019). Evaluating vacancy referrals and the roles of sanctions and sickness absence. *The Economic Journal*, *129*(624), 3292–3322.

Desiere, S., Langenbucher, K., & Struyven, L. (2019). *Statistical Profiling in Public Employment Services. OECD Social, Employment and Migration Working Papers, 224.*

Desiere, S., & Struyven, L. (2021). Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy*, *50*(2), 367–385.

DG EMPL (1990). *Employment in Europe.*

DG EMPL (2018). *Labour market policy statistics. Methodology 2018.* Publications Office of the European Union.

DG EMPL. (2020). *Labour market policy: Expenditure and participants : data 2018.* Publications Office.

DG EMPL. (2024). *LMP Expenditure by type of action—Summary tables (LMP_EXPSUMM)* [Dataset].

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, *27*(4), 605–615.

DiPrete, T. A., & Chae, J. (2023). Coordinated markets, school-to-work linkages, and labor market outcomes in Europe. *Research in Social Stratification and Mobility*, *87*, 100840.

DiPrete, T. A., Eller, C. C., Bol, T., & Werfhorst, H. G. (2017). School-to-work linkages in the United States, Germany, and France. *American Journal of Sociology*, *122*(6), 1869–1938.

Doerr, A., Fitzenberger, B., Kruppe, T., Paul, M., & Strittmatter, A. (2017). Employment and earnings effects of awarding training vouchers in Germany. *ILR Review*, *70*(3), 767–812.

Dong, Y., & Shen, S. (2018). Testing for rank invariance or similarity in program evaluation. *Review of Economics and Statistics*, *100*(1), 78-85.

Donoso, V., Martín, V., & Minondo, A. (2015). Do differences in the exposure to Chinese imports lead to differences in local labour market outcomes? An analysis for Spanish provinces. *Regional Studies*, *49*(10), 1746-1764.

Duell, N. H., & Moraes, G. (2023). *Statistical Profiling—Lessons from OECD Countries*. World Bank Group.

Dustmann, C., Lindner, A., Schönberg, U., Umkehrer, M., & Vom Berge, P. (2022). Reallocation effects of the minimum wage. *The Quarterly Journal of Economics*, *137*(1), 267–328.

Dye, T. R. (2012). *Understanding public policy* (14th ed.). Pearson.

Dyke, A., Heinrich, C. J., Mueser, P. R., Troske, K. R., & Jeon, K. (2006). The Effects of Welfare-to-Work Program Activities on Labor Market Outcomes. *Journal of Labor Economics*, *24*(3), 567–607.

Eggers, A. C., Tuñón, G., & Dafoe, A. (2024). Placebo Tests for Causal Inference. *American Journal of Political Science*, *68*(3), 1106–1121.

Eichhorst, W., & Marx, P. (2021). How stable is labour market dualism? Reforms of employment protection in nine European countries. *European Journal of Industrial Relations*, *27*(1), 93–110.

Eichler, M., & Lechner, M. (2002). An evaluation of public employment programmes in the East German State of Sachsen-Anhalt. *Labour Economics*, *9*(2), 143–186.

Engström, P., Hesselius, P., & Holmlund, B. (2012). Vacancy referrals, job search, and the duration of unemployment: A randomized experiment. *Labour*, *26*(4), 419–435.

Erixon, L. (2010). The Rehn-Meidner Model in Sweden: Its Rise, Challenges and Survival. *Journal of Economic Issues*, *44*(3), 677–715.

Escudero, V., Kluve, J., López Mourelo, E., & Pignatti, C. (2019). Active labour market programmes in Latin America and the Caribbean: Evidence from a meta-analysis. *The Journal of Development Studies*, *55*(12), 2644–2661.

Eurofound (2023). *Minimum wages: Non compliance and enforcement across EU Member States : comparative report.* Publications Office.

European Commission (2017). *Standard Eurobarometer 87—Spring 2017 [QA3]* [Dataset].

European Commission (2024). *Standard Eurobarometer 101—Spring 2024 [QA3]* [Dataset].

European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation. *Official Journal of the European Union*, *L 119*, 1–88.

Eurostat (2024a). *Annual national accounts* [Dataset].

Eurostat (2024b). *Demography, population stock and balance* [Dataset].

Eurostat (2024c). *Employment and unemployment (Labour force survey)* [Dataset].

Eurostat. (2024d). *Thematic glossaries*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Thematic_glossaries

Eurostat (2024e). *Total unemployment rate [tps00203].*

Eurostat (2024f). *Social protection and inclusion – headline indicators* [Dataset].

Everis. (2017). *Evaluación de implementación y de impacto de los Servicios de Orientación Profesional*. Servei Públic d'Ocupació de Catalunya.

Felgueroso, F., García-Pérez, J. I., Jiménez-Martín, S., Gorjón, L., & García, M. (2018). *Herramienta de perfilado de parados: Modelización y resultados preliminares* (F. Felgueroso, J. I. García-Pérez, & S. Jiménez-Martín, Eds.). Fundación Ramón Areces.

Fernández, R., Immervoll, H., Pacifico, D., Browne, J., Neumann, D., & Thévenot, C. (2018). Faces of Joblessness in Spain: A People-centred perspective on employment barriers and policies. *OECD Social, Employment and Migration Working Papers*, *207*.

Filges, T., Smedslund, G., & Jørgensen, A.-M. K. (2018). Active Labour Market Programme Participation for Unemployment Insurance Recipients: A Systematic Review. *Research on Social Work Practice*, *28*(6), 653–681.

Filomena, M. (2023). Unemployment scarring effects: An overview and meta-analysis of empirical studies. *Italian Economic Journal*, 1–60.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

Fitzenberger, B., Osikominu, A., & Paul, M. (2023). The effects of training incidence and planned training duration on labor market transitions. *Journal of Econometrics*, *235*(1), 256–279.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., & Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Review of Economics and Statistics*, *94*(1), 153–171.

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, *12*(1), 156–177.

Forster, A. G., & Bol, T. (2018). Vocational education and employment over the life course using a new measure of occupational specificity. *Social Science Research*, *70*, 176–197.

Fougère, D., Pradel, J., & Roger, M. (2009). Does the public employment service affect search effort and outcomes? *European Economic Review*, *53*(7), 846–869.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage publications.

Frantz, C. K., & Siddiki, S. (2022). *Institutional Grammar: Foundations and Applications for Institutional Analysis*. Springer International Publishing.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Frölich, M., & Sperlich, S. (2019). *Impact evaluation*. Cambridge University Press.

Gabrikova, B., Svabova, L., & Kramarova, K. (2023). Machine learning ensemble modelling for predicting unemployment duration. *Applied Sciences*, *13*(18), 10146.

Gallagher, P., & Griffin, R. (2023). *In) Accuracy in Algorithmic Profiling of the Unemployed–An Exploratory Review of Reporting Standards*. Social Policy and Society.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology*, *36*, 21–47.

García-Pérez, J. I., & Rebollo-Sanz, Y. F. (2009). The use of permanent contracts across Spanish regions: Do regional wage subsidies work? *Investigaciones Economicas*, *33*(1), 97–130.

Garnero, A., & Martin, S. (2023). *Collective bargaining coverage. A detailed methodological note*. OECD.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

González Veiga, M. C., Cueto Iglesias, B., & Mato Díaz, F. J. (2008). ¿Qué beneficios tiene para la empresa su colaboración con el sector educativo?: Un análisis de la Formación en Centros de Trabajo desde la perspectiva empresarial. *Revista del Ministerio de Trabajo y Asuntos Sociales: Revista del Ministerio de Trabajo e Inmigración*, *76*, 39–58.

Gorjón, L., Martinez De Lafuente, D., & Romero, G. (2024). Employment effects of the minimum wage: Evidence from the Spanish 2019 reform. *SERIEs*, *15*(1), 1–55.

Government of Catalonia. (2016). *ORDEN TSF/288/2016, de 24 de octubre*.

Government of Catalonia. (2017). *Resolució TSF/1897/2017, de 20 de juliol*.

Government of Veneto. (2017a). *Direttiva per la sperimentazione dell'Assegno per il lavoro per la ricollocazione di lavoratori disoccupati.*

Government of Veneto. (2017b). *Disposizioni in materia di tirocini ai sensi dell'art. 41 della legge regionale 13 marzo 2009: Vol. n. 3.*

Government of Veneto. (2017c). *Modello per la profilazione dei destinatari dell'Assegno per il lavoro (Allegato A al Decreto n.*

Government of Veneto. (2018). *Deliberazione della Giunta Regionale: Punto 73 dell'ODG della seduta del 30/04/2018.*

Grubb, D., and Puymoyen, A. (2008). *Long time-series for public expenditure on labour market programmes.* OECD - Directorate for employment, labour and social affairs.

Grubb, D., and Keese, M. (2018). *Methodological and operational issues encountered in OECD work with LMP data.* OECD.

Hanushek, E. A., Schwerdt, G., Woessmann, L., & Zhang, L. (2017). General education, vocational education, and labor-market outcomes over the lifecycle. *Journal of Human Resources*, *52*(1), 48–87.

Harding, D. J., & Seefeldt, K. S. (2013). Mixed Methods and Causal Analysis. In S. Morgan (Ed.), *Handbook of Causal Analysis for Social Research.* Springer.

Harmon, N. A., Mahlstedt, R., Rasmussen, M., & M. (2021). *Helping the Unemployed Through Statistical Prediction?* [Vocational training & statistical profiling [Doctoral thesis].].

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

Heckman, J. J., LaLonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1865–2097). Elsevier.

Hemerijck, A. (Ed.). (2017). *The Uses of Social Investment* (Vol. 1). Oxford University Press.

Hemerijck, A., Ronchi, S., & Plavgo, I. (2023). Social investment as a conceptual framework for analysing well-being returns and reforms in 21st century welfare states. *Socio-Economic Review*, *21*(1), 479-500.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62.

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, *26*(10), 674–680.

Hernán, M. A., & VanderWeele, T. J. (2011). *Compound Treatments and Transportability of Causal Inference*. Epidemiology.

Hernán, M. A., & Robins, J. M. (2023). *Causal Inference: What If*. Chapman & Hall/CRC.

Hick, R., & Marx, I. (2023). Poor workers in advanced democracies: On the nature of in-work poverty and its relationship to labour market policies. In D. Clegg & N. Durazzi (Eds.), *Handbook of Labour Market Policy in Advanced Democracies* (pp. 495-507). Edward Elgar Publishing.

Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, *13*(2).

Hirshleifer, S., McKenzie, D., Almeida, R., & Ridao-Cano, C. (2016). The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey. *The Economic Journal*, *126*(597), 2115–2146.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*(3), 199–236.

Howlett, M., & Goetz, K. H. (2014). Introduction: time, temporality and timescapes in administration and policy. *International Review of Administrative Sciences*, *80*(3), 477-492.

Howlett, M., & Mukherjee, I. (Eds.). (2018). *Routledge handbook of policy design*. Routledge.

Howlett, M. (2019). *The policy design primer: Choosing the right tools for the job*. Routledge.

Huber, M., Hsu, Y. C., Lee, Y. Y., & Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, *35*(7), 814–840.

Huber, M., Lechner, M., & Strittmatter, A. (2018). Direct and indirect effects of training vouchers for the unemployed. *Journal of the Royal Statistical Society. Series A (Statistics in Society*, *181*(2), 441–463.

Huuskonen, J. (2023). *Essays on the role of public employment services in labour market matching.* JYU Dissertations.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice. *OTexts*.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *171*(2), 481–502.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, *7*(1), 443–470.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G., & Xu, Y. (2024). *LaLonde (1986) after Nearly Four Decades: Lessons Learned*.

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, *58*(4), 1129–1179.

Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, *11*.

Immervoll, H., Knotz, C., & Otmani, I. (2020). *Activity-related elegibility conditions for receiving unemployment benefits. Results 2020*. OECD.

INE. (2024). *Contabilidad Nacional Trimestral de España. Primer trimestre 2024.* Avance.

Ivàlua. (2014). *Modelo de evaluación del SOC: conocimiento al servicio de las políticas activas de empleo.*

Jaenichen, U., & Stephan, G. (2011). The effectiveness of targeted wage subsidies for hard-to-place workers. *Applied Economics*, *43*(10), 1209–1225.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R.* Springer.

Jeremy, Z. (2023). *How Vacancy Referrals affect Job Search: Evidence from Job Applications Data.*

Junquera, Á. F. (2024a). *Efectos de la asistencia en la búsqueda de empleo: Una revisión sistemática para España, 35,* 7-25. Gestión y Análisis de Políticas Públicas.

Junquera, Á. F. (2024b). El diseño de los subsidios al empleo del Servei Públic d'Ocupació de Catalunya (2016-2018). In S. P. Guzmán & M. Iglesias (Eds.), *Precariedades laborales y desigualdades de género en Iberoamérica* (pp. 161–180). Dykinson.

Katz, L. F. (1998). Wage subsidies for the disadvantaged. In R. Freeman & P. Gottschalk (Eds.), *Generating jobs: How to increase demand for less-skilled workers.* Russell Sage Foundation.

Katz, L. F., Roth, J., Hendra, R., & Schaberg, K. (2022). Why do sectoral employment programs work? Lessons from WorkAdvance. *Journal of Labor Economics*, *40*(S1), 249–291.

Kemper, J., Stöterau, J., & Ghisletta, A. (2022). *The impact of vocational training interventions on youth labor market outcomes: A meta-analysis.*

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*(2), 131–159.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495.

Kluve, J. (2010). The effectiveness of European active labor market programs. *Labour economics*, *17*(6), 904–918.

Kluve, J., Puerto, S., Robalino, D., Romero, J. M., Rother, F., Stöterau, J., & Witte, M. (2019). Do youth employment programs improve labor market outcomes? A quantitative review. *World Development*, *114*, 237–253.

Koning, P., & Heinrich, C. J. (2013). Cream-skimming, parking and other intended and unintended effects of high-powered, performance-based contracts. *Journal of Policy Analysis and Management*, *32*(3), 461–483.

Körtner, J., & Bonoli, G. (2023). Predictive algorithms in the delivery of public employment services. In *Handbook of Labour Market Policy in Advanced Democracies* (pp. 387–398). Edward Elgar Publishing.

Kruppe, T., & Lang, J. (2018). Labour market effects of retraining for the unemployed: The role of occupations. *Applied Economics*, *50*(14), 1578–1600.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.

Kunaschk, M., & Lang, J. (2022). Can Algorithms Reliably Predict Long-Term Unemployment in Times of Crisis? – Evidence from the COVID-19 Pandemic. *IAB-Discussion Paper*.

Kuppler, M., Kern, C., Bach, R. L., & Kreuter, F. (2022). From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, *7*, 883999.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, *76*(4), 604–620.

Larsen, C. A., & Vesan, P. (2012). Why Public Employment Services Always Fail. Double-Sided Asymmetric Information and the Placement of Low-Skill Workers in six European Countries. *Public Administration*, *90*(2), 466–479.

Lasswell, H. D. (1951). The policy orientation. In D. Lerner & H. D. Lasswell (Eds.), *The policy sciences: Recent developments in scope and method* (pp. 3–15). Stanford University Press.

Layard, P. R. G., Nickell, S. J., & Jackman, R. (2005). *Unemployment: Macroeconomic performance and the labour market* (2nd ed., new ed.). Oxford University Press.

Lechner, M. (2000). An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany. *The Journal of Human Resources*, *35*(2), 347.

Lechner, M., Miquel, R., & Wunsch, C. (2007). The curse and blessing of training the unemployed in a changing economy: The case of East Germany after unification. *German Economic Review*, *8*(4), 468–509.

Lechner, M., & Wunsch, C. (2009). Are training programs more effective when unemployment is high? *Journal of Labor Economics*, *27*(4), 653–692.

Lee, W. (2009). Private Deception and the Rise of Public Employment Offices in the United States, 1890– 1930. In D. H. Autor (Ed.), *Studies of Labor Market Intermediation* (pp. 155–182). University of Chicago Press.

Leuven, E. (2005). The economics of private sector training: A survey of the literature. *Journal of Economic Surveys*, *19*(1), 91–111.

Levy Yeyati, E., Montané, M., & Sartorio, L. (2019). *What works for active labor market policies? CID Working Paper Series.*

Linares Martínez, F. (2018). *Sociología y teoría social analíticas: La ciencia de las consecuencias inintencionadas de la acción.* Alianza Editorial.

Linfante, G., Radicchia, D., Stocco, P., & Toti, E. (2019). Rapporto di valutazione della sperimentazione dell'assegno di ricollocazione. *ANPAL*, *8.*

Liu, S., Huang, J. L., & Wang, M. (2014). Effectiveness of job search interventions: A meta-analytic review. *Psychological Bulletin*, *140*(4), 1009.

Loxha, A., & Morgandi, M. (2014). Profiling the unemployed: A review of OECD experiences and implications for emerging economies. In *Social Protection and labor discussion paper, SP 1424.*

Maibom, J., Rosholm, M., & Svarer, M. (2017). Experimental evidence on the effects of early meetings and activation. *The Scandinavian Journal of Economics*, *119*(3), 541–570.

Malani, A., & Reif, J. (2015). Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics*, *124*, 1–17.

Manski, C. F., & Pepper, J. V. (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, *100*(2), 232–244.

Margaryan, S., Saniter, N., Schumann, M., & Siedler, T. (2022). Do internships pay off? The effects of student internships on earnings. *Journal of Human Resources*, *57*(4), 1242–1275.

Martin, J. P. (2015). Activation and active labour market policies in OECD countries: Stylised facts and evidence on their effectiveness. *IZA Journal of Labor Policy*, *4*, 1–29.

Martinson, K., Harvill, E., Litwok, D., Schwartz, D., Rosa, S. M., Saunders, C., & Bell, S. (2019). Implementation and Relative Impacts of Two Job Search Assistance Programs in New York City. *The Job Search Assistance Strategies Evaluation. Office of Planning.*

Marx, C., Calmon, F., & Ustun, B. (2020). Predictive multiplicity in classification. *International Conference on Machine Learning*, 6765–6774.

Mato-Díaz, F. J., & Escudero-Castillo, I. (2023). Lost in transition? Young graduates' employment after internships in public and private organisations. *Journal of Youth Studies*, 1–17.

Matty, S. (2013). *Predicting likelihood of long-term unemployment: The development of a UK jobseekers' classification instrument.* Corporate Document Services.

McGuinness, S., O'Connell, P. J., & Kelly, E. (2014). The impact of training programme type and duration on the employment chances of the unemployed in Ireland. *The Economic and Social Review*, *45(3*, 425–450.

McGuinness, S., Redmond, P., Kelly, E., & Maragkou, K. (2022). *Predicting the probability of long-term unemployment and recalibrating Ireland's statistical profiling model* (Research Series, Issue 149)).

Meidner, R. (1969). Active Manpower Policy and the Inflation Unemployment-Dilemma. *The Swedish Journal of Economics*, *71*(3), 161.

Meyer, B. D. (1995). Lessons from the US unemployment insurance experiments. *Journal of Economic Literature*, *33*(1), 91–131.

Miguel, P., Serra, J. A., & Caballero, M. (2021). *INVOLVE Project: Desk research Spain.*

Ministry of Labor of the Government of Spain. (2024). *Anuarios de Estadísticas del Ministerio* [Dataset].

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*(1), 141–163.

Mohan, K., & Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, *116*(534), 1023–1037.

Molina, O., & Rhodes, M. (2007). The political economy of adjustment in mixed market economies: A study of Spain and Italy. In R. Hancké, M. Rhodes, & M. Thatcher (Eds.), *Beyond varieties of capitalism*, 223-252. Oxford University Press.

Molina, O., Junquera, Á. F., Verd Pericàs, J. M., Sánchez Martínez, R., Úbeda Molla, P., Galobardes, M., & Miró Martín, S. (2023). *PerfilaSP - Anàlisi de variables sociològiques i psicològiques com a predictores d'ocupabilitat en el perfilatge dels Serveis d'Ocupació de Catalunya.*

Molnar, C. (2022). *Interpretable machine learning. A Guide for Making Black Box Models Explainable.*

Muñiz, F. (2021). *Send@: Digitalización y uso masivo de datos para ayudar a encontrar trabajo—SEPE x JOBMadrid'20.* YouTube. https://www.youtube.com/watch?v=TNKVWL0pFRU

Nedelkoska, L., & Quintini, G. (2018). *Automation, skills use and training.* OECD Social, Employment and Migration Working Papers No. 202.

Nelson, K., Nieuwenhuis, R., & Yerkes, M. A. (2022). *Social policy in changing European societies.* Edward Elgar Publishing.

OECD (1990). *Labour Market Policies for the 1990s.* OECD.

OECD (2006). *OECD Employment Outlook 2006: Boosting jobs and incomes.* OECD.

OECD (2018). *OECD Handbook for Internationally Comparative Education Statistics 2018: Concepts, Standards, Definitions and Classifications.* OECD.

OECD (2019a). *OECD Employment Outlook 2019: The future of work.* OECD.

OECD (2019b). *Strengthening Active Labour Market Policies in Italy.* OECD.

OECD (2020). *OECD Employment Outlook 2020: Worker Security and the COVID-19 Crisis.* OECD.

OECD (2021). *The New OECD Employment Protection Legislation Indicators for Temporary Contracts.* OECD.

OECD (2022). *Coverage and classification of OECD data for public expenditure and participants in labour market programmes.*

OECD (2023a). *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market.* OECD.

OECD (2023). *Boosting Social Inclusion in Spain: Improving Pathways and Coordination of Services.* OECD. https://doi.org/10.1787/56b604a0-en.

OECD. (2023c). *Education at a Glance 2023: OECD Indicators.* OECD.

OECD (2023d). *OECD/AIAS ICTWSS database version 1.1 (September 2023)* [Dataset].

OECD (2024). *Indicator: Benefits in unemployment, share of previous income.* [Dataset].

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, *16*, 1–21.

Orfao, G., & Malo, M. Á. (2023). Are active labour market policies effective for the older unemployed? A meta-evaluation. *Ageing & Society*, *43*(7), 1617–1637.

Osikominu, A. (2013). Quick job entry or long-term human capital development? The dynamic effects of alternative training schemes. *Review of Economic Studies*, *80*(1), 313–342.

Pacifico, D., Rastrigina, O., Bachelet, M., Browne, J., Immervoll, H., Neumann, D., Tenaglia, S., Nicola, F., & Cozzolino, M. (2018). *The OECD Tax-Benefit model for Italy. Description of policy rules for 2018*. OECD – Directorate for Employment, Labour, and Social Affairs.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4), 579–595.

Pei, Z., Lee, D. S., Card, D., & Weber, A. (2022). Local polynomial order in regression discontinuity designs. *Journal of Business & Economic Statistics*, *40*(3), 1259–1267.

Picchio, M., & Ubaldi, M. (2022). Unemployment and health: A meta-analysis. *Journal of Economic Surveys*.

Pérez del Prado, D. (2021). *El impacto social de la Gobernanza económica europea*. Tirant lo Blanch.

Picot, G. (2023). Minimum wages: By collective bargaining and by law. In D. Clegg & N. Durazzi (Eds.), *Handbook of Labour Market Policy in Advanced Democracies* (pp. 234–248). Edward Elgar Publishing.

Polavieja, J. G., Lancee, B., Ramos, M., Veit, S., & Yemane, R. (2023). In your face: A comparative field experiment on racial discrimination in Europe. *Socio-Economic Review*, *21*(3), 1551–1578.

Potter, S., Enoch, M., Rye, T., Black, C., & Ubbels, B. (2006). Tax Treatment of Employer Commuting Support: An International Review. *Transport Reviews*, *26*(2), 221–237.

Qu, Z., & Yoon, J. (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics*, *185*(1), 1–19.

Qu, Z., & Yoon, J. (2019). Uniform inference on quantile effects under sharp regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(4), 625–647.

Rambachan, A., & Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, *90*(5), 2555–2591.

Rebollo-Sanz, Y. F. (2018). *El modelo de perfilado estadístico: Una herramienta eficiente para caracterizar a los demandantes de empleo* (F. Felgueroso, J. I. García-Pérez, & S. Jiménez-Martín, Eds.). Fundación Ramón Areces.

Rehn, G. (1985). Swedish active labor market policy: Retrospect and prospect. *Industrial Relations: A Journal of Economy and Society*, *24*(1), 62–89.

Rehn, G. (1989). *Ensayos sobre política de empleo activa*. Ministerio de Trabajo y Seguridad Social.

Rehwald, K., Rosholm, M., & Svarer, M. (2017). Do public or private providers of employment services matter for employment? Evidence from a randomized experiment. *Labour Economics*, *45*, 169–187.

Rocha, F. (2018). Strenthening the decentralisation of collective bargaining in Spain. Between the legal changes and the real developments. In S. Leonardi & R. Pedersini (Eds.), *Multi-employer bargaining under pressure: decentralisation trends in five European countries*. European Trade Union Institute.

Rojas, P. N. (2012). La dimensión local de las políticas activas de empleo. Una revisión competencial tras la aprobación de la estrategia española de empleo 2012-2014. *Temas Laborales: Revista Andaluza de Trabajo y Bienestar Social*, *114*, 119–146.

Rojo Torrecilla, E. (2022). La reforma laboral de 2021. Negociación colectiva, flexibilidad interna y subcontratación. *Revista de Trabajo y Seguridad Social. CEF*, 77–117.

Román, C., Congregado, E., & Millán, J. M. (2013). Start-up incentives: Entrepreneurship policy or active labour market programme? *Journal of Business Venturing*, *28*(1), 151–175.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer.

Rözer, J. J., & Bol, T. (2019). Labour market effects of general and vocational education over the life-cycle and across time: Accounting for age, period, and cohort effects. *European Sociological Review*, *35*(5), 701–717.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36.

Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 808–840.

Sabatier, P., & Mazmanian, D. (1980). The implementation of public policy: A framework of analysis. *Policy studies journal, 8*(4), 538-560.

Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, *219*(1), 101–122.

Scarano, G. (2019). *Towards a new public employment service? Finding the optimal design of public-private contracting arrangements. Evidence from two regional cases* [Universit degli Studi di Milano].

Scarano, G. (2023). Outcome-based contracting and gaming practices in marketised public employment services. Dilemmas from the Italian case. *Journal of Social Policy*, 1–19.

Schneider, A. and Ingram, H. (1997). *Policy Design for Democracy.* University of Kansas Press.

Siddiki, S., & Curley, C. (2022). Conceptualising policy design in the policy process. *Policy & Politics*, *50*(1), 117-135.

SOC (2016a). *Manual del model operatiu de les oficines de treball.* Confidential document.

SOC (2016b). *Resolució TSF/2392/2016, de 25 d'octubre.*

SOC (2023a). *Especialitats formatives i unitats formatives del Servei Públic d'Ocupació.*

SOC (2023b). *Habilitacions professionals dels Certificats de Professionalitat.*

SOC (2023c). *Pla d'acció per a persones en situació d'atur majors de 45 anys i de llarga durada 2023-2026.*

SOC (2024). *Històric de dades d'inversió* [Dataset].

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science*, *25*(1), 1-21.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), 84–90.

Subirats, J., Knoepfel, P., Larrue, C., & Varone. (2008). *Análisis y gestión de políticas públicas* (1. ed). Ariel.

Titiunik, R. (2021). Natural Experiments. In J. N. Druckman & D. P. Gree (Eds.), *Advances in Experimental Political Science* (pp. 103–129). Cambridge University Press.

Troya, I. M., Chen, R., Moraes, L. O., Bajaj, P., Kupersmith, J., Ghani, R., & Zejnilovic, L. (2018). Predicting, explaining, and understanding risk of long-term unemployment. *32nd Conference on Neural Information Processing Systems.*

UNESCO. (2012). *International standard classification of education: ISCED 2011.* UNESCO Institute for Statistics.

van Dijk, W. (2019). *The socio-economic consequences of housing assistance.*

Viljanen, M., & Pahikkala, T. (2020). Predicting unemployment with machine learning based on registry data. In F. Dalpiaz, J. Zdravkovic, & P. Loucopoulos (Eds.), *Research Challenges in Information Science* (pp. 352–368). Springer International Publishing.

Visser, J. (2016). What happened to collective bargaining during the great recession? *IZA Journal of Labor Policy*, *5*(1), 9.

Visser, J. (2021). *OECD/AIAS ICTWSS Database. Detailed note on definitions, measurement and sources.* OECD.

Visser, J. (2024). Did employers abandon collective bargaining? A comparative analysis of the weakening of collective bargaining in the OECD. *Industrial Relations Journal*, *55*(5), 350–377.

Vooren, M., Haelermans, C., Groot, W., & Brink, H. (2019). The effectiveness of active labor market policies: A meta-analysis. *Journal of Economic Surveys*, *33*(1), 125–149.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Weishaupt, J. T. (2011). *From the Manpower Revolution to the Activation Paradigm: Explaining Institutional Continuity and Change in an Integrating Europe*. Amsterdam University Press.

Winterhager, H., Heinze, A., & Spermann, A. (2006). Deregulating job placement in Europe: A microeconometric evaluation of an innovative voucher scheme in Germany. *Labour Economics*, *13*(4), 505–517.

Wood, J., Neels, K., & Vujić, S. (2024). Which training leads to employment? The effectiveness of varying types of training programmes for unemployed jobseekers in Flanders. *Journal of Social Policy*, 1–22.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, *26*(3), 31–66.

World Bank (2001). *Social protection sector strategy: From safety net to springboard*. The World Bank.

Wu, C. J., & Hamada, M. S. (2021). *Experiments: Planning, analysis, and optimization*. John Wiley & Sons.

Xu, K. L. (2017). Regression discontinuity with categorical outcomes. *Journal of Econometrics*, *201*(1), 1–18.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, *25*(1), 57–76.

Yang, S., Lorenzi, E., Papadogeorgou, G., Wojdyla, D. M., Li, F., & Thomas, L. E. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine*, *40*(19), 4294–4309.

Yu, R., & Ding, P. (2023). *Balancing weights for causal inference in observational factorial studies.*

Zanella, G., & Salomone, R. (2024). *Incentive-based active labor market programs: Insights from two policy experiments in Italy.*

Zeldow, B., & Hatfield, L. A. (2021). Confounding and regression adjustment in difference-in-differences studies. *Health Services Research*, *56*(5), 932–941.

Zuchuat, J. (2023). *How Vacancy Referrals affect Job Search: Evidence from Job Applications Data.*