Performance Pay for Private Program Providers and Impact on Participants: A Field Experiment with Employment Services in Norway[*]

Øystein M. Hernæs, Ragnar Frisch Centre for Economic Research

June 17, 2025

Abstract:

This study examines whether performance-based pay for private employment service providers improves employment outcomes for program participants compared to traditional hourly compensation. Finding effective ways to outsource public services to external providers has the potential to improve the quality and efficiency of these services. Using a large-scale randomized controlled trial (RCT) conducted in Norway from April 2018 to December 2021, we evaluate the impact of different payment models on the performance of private firms in delivering labor market programs. A total of 4,898 unemployed individuals were randomly assigned to either a treatment group (37%), where providers received performance-based pay contingent on participants' employment outcomes, or a control group (63%), where providers were compensated on an hourly basis. Despite the substantial financial incentives involved, our findings reveal no significant differences in employment rates, earnings, or hours worked between the two groups. The results allow us to rule out effects on monthly earnings of ± €5 and employment effects of ± 1 percentage points after 12 months. There were no indications of heterogeneous treatment effects across different participant groups. A cost-benefit analysis suggests a supportive case for performance pay due to lower public costs, although this estimate is subject to uncertainty.

JEL classification: J64, J68, L33, H43

Keywords: Active labor market policies, Contracting out, Job-search assistance, Performance pay, Private provision of employment services

[*] The research has been funded by the Norwegian Labour and Welfare Administration (NAV). This paper builds on the evaluation conducted by Helene Berg and Gro Malene Vestøl (Proba samfunnsanalyse) and Øystein M. Hernæs (Frisch Centre), see Proba (2023) for details. Lars Bakken (NAV) was invaluable in ensuring the successful execution of the project. Data made available by NAV has been essential for the project. The empirical analysis was pre-registered with the Open Science Framework (OSF) at https://doi.org/10.17605/OSF.IO/HDJ6M. I thank three anonymous referees for a number of constructive comments.

1 Introduction

Finding the most effective way to outsource public services to external providers has the potential to improve the quality and efficiency of these services. In the context of employment programs, performance-based pay has been proposed as a way to better align providers' incentives with policy goals. However, empirical evidence on its effectiveness remains limited. This study tests whether performance-based pay for private employment service providers improves employment outcomes compared to traditional hourly compensation through a field experiement conducted in Norway from April 2018 to December 2021.

Contracting out has long been viewed as a way of reforming public services to capture some of the benefits of private-sector incentives for quality improvement and cost reduction without full privatization (Hart et al., 1997; Blank, 2000). Early empirical work reported substantial savings (Domberger and Jensen 1997), but later reviews emphasize that benefits depend on how easily quality can be contracted. Andersson et al. (2019) reviewed the literature on public service outsourcing and found positive results of private provision for services without severe contracting problems, but mixed evidence for services with severe contracting problems, such as employment services.

Crépon (2018) reviewed the evidence on private provision of labor market services and concluded that private provision seems to make relatively little difference for transitions to work. Three previous studies from Scandinavia, Bennmarker et al. (2013), Laun and Thoursie (2014) and Rehwald, Rosholm, and Svarer (2017), likewise found little difference in employment outcomes despite private providers using markedly different approaches. More recently, Egebark et al. (2024) analyzed Sweden's new "Prepare-and-Match" scheme, a support program for the long-term unemployed with a medium level of need for support, delivered by independent contractors with greater freedom in what to offer. The evaluation, which was based on a randomized controlled trial, showed that the new program increased the amount (and cost) of support, but did not raise earnings, employment, or training participation.

Because contracting on service quality or content is difficult, there has been interest in tying at least part of providers' compensation directly to measurable performance. However, the existing literature on public sector outsourcing offers mixed results regarding the effectiveness of such performance-based financing of employment services (Langenbucher and Vodopivec, 2022; Stephan, 2024). Stephan (2024) also discussed conditions for successful contracting out job placement services, primarily including sufficiently complete contracts and adequate monitoring of quantity and quality. Results-based financing schemes attempt to solve the contracting problem by contracting directly on employment, the central goal of most labor market programs. While performance-based financing may enhance quality and reduce costs, concerns arise about potential unintended consequences like selectively choosing participants most likely to trigger results-based payments ("cream-skimming"), allocating minimal resources to participants unlikely to generate payments ("parking"), or the prioritization of short-term employment outcomes over long-term job retention and career progression.

A small number of studies investigate the impact of results-based pay in the context of employment programs. Koning and Heinrich (2013), using a difference-in-differences design, found that moving from partial to full performance-contingent pay to private providers in the Netherlands had a positive effect on short-term job placement for more readily employable workers, but also that it led to some cream-skimming. Koning and van de Meerendonk (2014) found that higher weight on providers' reputation and methodology in the tender evaluation process increased job placement in the Netherlands. Homrighausen (2014), comparing the employment outcomes of participants served by providers with different contract types under a selection-on-observables assumption, found that high performance-based payments were associated with increased job retention, while high upfront payment were associated with lower job retention. Egemark et al. (2024) primarily evaluated the effect of private provision, but also compared estimated impacts for people randomized to providers with different payment schemes. Paying more per participant did not improve outcomes, but the evaluation found suggestive evidence that providers with a higher degree of results-based financing had somewhat better effects on employment.

This paper contributes to this literature by offering credible empirical evidence on the impact of payment schemes on employment service delivery from a large-scale, individual-level randomized controlled trial. Specifically, we ask: Does a performance-based pay scheme improve employment outcomes for unemployed individuals compared to the standard hourly pay model? Randomizing participants to providers with different pay structures isolates the effect of incentive structures on participants' employment outcomes. The performance pay arm involved contracting on both obtaining and retaining employment. This compares to the control arm, which did not have any results-based component. Both schemes required monitoring, in the results-based arm to verify that placement and retention criteria were met and that no participants were neglected ("parked"), while in the hourly-fee scheme involved regular monitoring of the content and delivery of counselling sessions to ensure service quality.

We can rule out effects on monthly earnings of ± €5 and employment effects of ± 1 percentage points for participants allocated to a provider with performance pay. Moreover, there was no evidence of heterogeneous treatment effects across different participant groups. A cost-benefit analysis suggests a supportive case for performance pay due to lower public costs of around €1600 per program participant, although this estimate is subject to considerable uncertainty. One possible explanation, advocated by several of the involved practitioners, for why performance pay didn't make a substantial difference is that the comparison program already delivered high-quality, intensive support, thus setting a high bar for any additional impact.

The rest of the paper is structured as follows: Section 2 provides institutional background, section 3 describes the experimental design and methodology for the field experiment, section 4 presents the data, and section 5 presents the results. Section 6 concludes and discusses the implications of the findings for policy and practice.

2 Institutional setting

The Norwegian Labour and Welfare Administration (NAV) is responsible for administering both labor market programs. Historically, the public employment services in Norway have been relatively active, with a large share of the unemployed participating in such programs. Most programs are work-oriented, focusing on rapid transitions to employment through various forms of support, work practice, or wage subsidies. In addition, there are qualification programs aimed at enhancing skills or providing formal credentials, as well as assessment programs designed to evaluate participants' work capacity and identify appropriate next steps.

All labor market programs are publicly funded, but private providers play an important role. In particular, many of the work-oriented programs are procured from external providers. These programs are acquired through competitive tendering and typically financed via an hourly payment model. This payment model compensates providers based on the accrued hours spent on direct participant follow-up, with a performance target set as the basis for evaluating, though not paying, program suppliers.

Performance-based financing models for work-oriented programs have been tested intermittently in Norway. Initial efforts in 2002 involved bonus financing in job provision assistance, job clubs, and a program for individuals receiving a so-called "waiting benefit".[1] Lunder et al. (2005) found no significant effects of the bonus scheme, possibly due to the small size of bonus payments. Between 2008 and 2011, a new trial introduced two alternative financing models for the "Work with Assistance" program. Econ and Proba (2010) found no effects on the transition to work, but the study was limited by a small degree of results-based funding and a low number of survey respondents.

In 2012, the government-appointed Brofoss Commission (NOU 2012:6) recommended evaluating alternative financing forms, including results-based models. Interest in performance-based financing was related to the international concern with the effectiveness of traditional input- or activity-based contracts that paid for participation rather than job outcomes. As noted in NOU 2012:6, results-based models aimed to make services more relevant to individual job seekers, improve quality and strengthen incentives to move participants into regular work. Input-based contracts were criticized for rewarding service volume rather than effectiveness and, in some cases, for disincentivizing placements if these reduced billable activities. The Commission recommended piloting milestone- and outcome-based financing.

Following up on this recommendation, a pilot was launched in 2013 in which a larger share of the payment depended on results. Proba (2014) evaluated the implementation and experiences with the first phase of this pilot, but the study was constrained by a limited number of participants and reliance on individual NAV supervisors for recruitment.

[1] Original program names: Formidlingstiltak (KAT), jobbklubb (JmB), ventestønad (VALS).

The government white paper Meld. St. 33 (2015-2016) stated that it was a political goal to continue to study incentive-based systems, including labor market programs that incorporated the use of financial incentives into provider contracts. Consequently, the Ministry of Labour and Welfare initiated a new trial with results-based financing of program providers in order to obtain better evidence, this time as a randomized controlled trial with a large sample. This trial provides an opportunity to rigorously investigate the effects of the payment model on employment outcomes.

3 The field experiment

3.1 Setting: The follow-up program

The setting of the study was the "follow-up program" ("Oppfølgingstiltaket"), which is the largest labor market program in Norway. The program targets individuals who require support to secure and sustain employment. It aims to provide tailored assistance to each participant, guiding them towards achieving and sustaining paid work. The program encompasses a suite of services, including a thorough needs assessment, job search support, training in social and work-related skills, and ongoing follow-up of both the participant and, when relevant, the employer. Each participant has one specific "job specialist" responsible for all program services related to that participant. This comprehensive model is designed to address the multifaceted challenges faced by individuals on the margins of the labor market and aim to bridge the gap between unemployment and stable employment.

Participants in the control group were randomized to the standard follow-up program, in which providers were compensated on an hourly basis. This payment model reimbursed providers for time spent supporting participants and was governed by a detailed requirements specification outlining provider obligations. The requirement specification included the following: "NAV assumes an average need of 10 hours of follow-up per participant per month throughout the duration of the program. However, this may vary from one participant to another, and the number of hours used should be adjusted to individual needs." In interviews, job specialists reported that even though 10 hours was specified as an average, most participants in practice received 10 hours per month, regardless of individual needs.

In contrast, the treatment group operated under a performance pay model. These providers were compensated based on the results they achieved, namely the participants' job placement and job retention. Accompanying this payment model was a less detailed requirement specification, allowing providers greater flexibility in service delivery. The statement about hours was changed to: "Participants shall receive the number of follow-up hours deemed necessary to achieve the goal of entering and/or maintaining employment. The need for follow-up hours may vary from one participant to another, and the number of hours used should be tailored to individual needs." According to job specialists, this formulation allowed them to be more flexible and prioritize their time where it was most needed.

Providers delivering both programs were subject to the standard monitoring and quality assurance procedures established for NAV-contracted services. These included the use of NAV's Quality Evaluation Tool, which involves systematic review of provider practices based on the specific contract text. NAV also generally monitors results through its data warehouse systems and receives regular reports from providers. Any specific contractual requirements, such as the obligation to use separate premises for different programs in this case, are supposed to be followed up explicitly. In general, NAV has a formal obligation to oversee the quality, outcomes, and content of contracted services. This responsibility is also reflected in NAV's annual objectives and allocation letter to the county offices, which include guidance on monitoring procedures and expectations for follow-up. Additionally, guidance from Anskaffelser.no and the Norwegian Agency for Public and Financial Management (DFØ) outlines how public procurement contracts should be followed up, including expectations for contract management.

3.2 The performance pay contract

The performance pay contract consisted of three milestones: An assessment milestone, an obtaining employment milestone and a retaining employment milestone. Each is described in turn below.

Introductory phase (20% of total payment): Providers received 20% of the agreed price for each participant who completed an initial phase lasting up to 2 months. This phase involved an investigation of the participant's needs and experiences, and the development of a personalized plan for work-related activity.

The remaining 80% of the payment was contingent on the participant securing and retaining regular employment. 40% was linked to obtaining regular employment, while the final 40% was linked to retaining employment over time. The rules differd depending on whether the job was temporary or permanent. The employment milestones were defined as follows:

Obtaining employment (40%): Permanent employment in one position in at least 50% of full time. For temporary employment, the required work time was also at least 50% of full time, but the position had to have a continuous duration of more than six months. In the participant held multiple jobs, at least one job had to meet the 50% threshold. For participants on sick leave from an ongoing work-relationship at the start of the program, the obtaining employment-condition was fulfilled if the person achieved (physician-certified) 80% recovery and had a plan for 100% recovery. For participants receiving disability benefits at the start of the program, the condition was fulfilled if the disability benefit was reduced, irrespective of the work fraction.

Retaining employment (40%): Permanent employment with a continuous duration of more than six months. For temporary employment, employment had to have a continuous duration of more than nine months (three additional months beyond the threshold for initial job obtainment). For participants on sick leave at the start of the program, the condition was fulfilled if the participant had fully recovered and ramined employed for more than six months.

This payment structure was intended to incentivize providers to both place participants in jobs and support them in retaining those jobs over time. All milestones had to be achieved within ten months after program completion.

3.3 Choice of program providers

Due to legal constraints under Norwegian procurement law, program providers in the study had to be selected through competitive public tenders in each region. Price, defined as the total potential compensation per participant, was one of the key criteria in the tendering process. In addition to price, the tenders included quality-related criteria. Specifically, providers were required to describe in detail how they would implement and quality-assure the intervention. This included describing planned activities and themes, how the employment focus would be maintained, and providing a credible explanation of how the proposed approach would meet the needs of the target group. Providers also had to specify the work processes, tools, and methods they intended to use. The weighting of price and quality criteria varied by county. In the two cases where we were able to obtain detailed evaluation documentation, price was weighted at 40% and 50%, while the remaining 60% and 50% were allocated to various dimensions of service quality.

Because the program providers were selected in this way, there was no randomization of providers to financing schemes. This raises the possibility that providers may have selected into schemes that better matched their preferences or competencies. If so, differences in outcomes could reflect differences in provider characteristics rather than the payment model alone.

To assess this possibility, Table 1 presents an overview of the providers in our setting. In each of the five regions under study, one provider was selected through the tendering process to deliver the program with performance pay. We denote the treatment providers A, B, C and D, and the control providers J, D, A, K, B and A. To be clear, the later empirical analysis will estimate a treatment effect that is common for all treatment providers, evaluated against all control providers.

Panel A compares important observable characteristics across treatment and control providers. We can see that providers varied considerably in size, though this variation was present also within the treatment and control groups. Firm A, a large for-profit provider, and firm D, a large non-profit were notably larger than the other firms. Both these firms operated under both schemes in different counties. A small firm, firm B, also also participated in both models. Most of the providers in both groups were for-profit firms. The performance pay group included for-profit providers and one non-profit provider, while the hourly pay group included three for-profit providers, one non-profit, and one area divided between a for-profit and a municipal provider. Firm size, measured by revenue and employment (in man-years), was not systematically larger or smaller in one scheme relative to the other. Overall, the distribution of provider characteristics suggests that selection into contract types did not produce large observable differences in provider profiles.

Panel B provides information about the runner-up bidders for the performance pay program and any changes in the provider of the control program during the study period. We first note that there

was substantial overlap of bidders across counties for the performance pay contracts. Of the four firms that ultimately received performance pay contracts – firms A, B, C and D, three – firms A, B, and D – also delivered the regular hourly pay program in at least one region. This overlap reflects the relatively thin market for these services in Norway, with both treatment and control programs provided by a limited set of firms. This also suggests that differences in provider characteristics are less likely to drive the observed outcomes, supporting the internal validity of the comparison between payment schemes.

Panel A

Performance pay (treatment)				Hourly pay (control)			
Provider	Revenue	Employed	Ownership	Provider	Revenue	Employed	Ownership
A	25,321	262	FP	J	25,321	29	FP
A	25,321	262	FP	D	26,789	259	NP
B	3,061	12	FP	A	25,321	262	FP
C	97	4	FP	K / B	unknown/3,061	unknown/12	M / FP
D	26,789	259	NP	A	25,321	262	FP

Panel B

County	Region	Provider with performance pay	Provider with hourly pay	Change of provider with hourly pay
Akershus	Ullensaker and Eidsvoll	A (Runner-up: D, C, E)	J	No change
Hordaland	Bergen and Nord-Hordaland	A (Runner-up: C, D, F)	D	A (August 2020)
Oslo	Vest	B (Runner-up: C, A, D)	A	No change
Telemark	Grenland and Vestmar	C (Runner-up: G, H, E)	Region 1: K Region 2: B	L (July 2020)
Østfold	Sarpsborg	D (Runner-up: I, C, A)	A	J (July 2019)

Table 1. Characteristics of the program providers

Note: Revenue in 2018, measured in 1,000 €. Employed in 2018, measured in man-years. Ownership: FP: For-profit (commercial), NP: Non-profit, M: Municipal business enterprise. Source: The Brønnøysund Register Centre.

Another potential concern is that firms delivering both performance pay and hourly pay programs might have strategically prioritized one scheme over the other. To mitigate this risk, the requirements specification for the performance pay contracts explicitly mandated that, in cases where a provider operated both programs, staff and physical locations had to be kept entirely separate. These provisions aimed to prevent both deliberate resource reallocation and unintended spillovers that could compromise the internal validity of the study. Thus, while we cannot rule out such behavior, contractual safeguards were explicitly designed to limit cross-contamination and strategic prioritization.

3.4 Randomization

The experiment took place in five of Norway's 18 counties. In each county, a designated NAV staff member was responsible for the randomization procedure. Whenever a jobseeker at the local social security administration in one of the trial areas was deemed eligible for the follow-up program, the person's caseworker phoned the person responsible for randomization in the county. This person logged into a decicated website we had set up, entered the job-seeker's identification number, after which a randomization algorithm assigned the jobseeker to the treatment or control group. The computer generated a pseudo-random number uniformly distributed between 0 and 1. If this number was lower than a specified threshold, the person was allocated to the treatment group.

From April 2018 to December 2021, 4,898 unemployed individuals, deemed eligible for the follow-up program by their local employment office, were randomly assigned to either the treatment group (37%), where program providers received performance-based pay contingent on participants' employment outcomes, or the control group (63%), where providers were compensated on an hourly basis. Figure A1 in the appendix shows the number of individuals randomized to the two conditions by month.

The treatment group was randomized to program providers that were financed based on the participants' employment performance. The control group consisted of program participants randomized to providers financed by the pay-per-hour-model. Participants were individually randomized to treatment or control within each county, thus treatment and control groups should (in expectation) be balanced within each county. Not all counties had the same share of treatment vs. control group individuals. As there are demographic differences between the counties, the sample was not necessarily balanced in the aggregate.

There was no blinding in the study. The program providers were fully informed about the details of the experiment, while the program participants may or may not have been aware of the assigned treatments. The personnel who interacted directly with the study subjects were aware of the assigned treatments, but were contractually obliged to work with only one of the groups.

The trial covered five counties, Akershus, Hordaland, Oslo, Telemark and Østfold, and ran from April 2018 to December 2021. In this period, 4898 individuals. were randomized, of which 1825 to the treatment group and 3073 to the control group.

The main hypothesis was that the treatment group would have better employment outcomes because of the stronger incentives of their program providers. Employment outcomes were measured using automatically collected administrative data.

4 Data

The study is based on administrative register data provided by Norwegian Labour and Welfare Administration.[2] The data cover year of birth, gender, immigrant status, income from various sources, employment history, monthly employment records, program participation, and welfare benefit claims.

Table 2 presents descriptive statistics on gender, an indicator variable of whether the person is an immigrant, age, labor market experience (total number of years with positive labor market earnings), labor market earnings in the previous year and total income in the previous year. The table shows means and standard deviations in both groups and tests differences using t-tests. We see that the treatment group is on average 0.76 years younger (significant at the 5% level) and has slightly fewer years of work experience and higher prior earnings.

Table 2. Descriptive statistics.

	All		Treatment		Control		Difference	P-value
	mean	sd	mean	sd	mean	sd	(C-T)	
Female	0.54	(0.50)	0.55	(0.50)	0.54	(0.50)	-0.01	0.662
Immigrant	0.28	(0.45)	0.28	(0.45)	0.29	(0.45)	0.00	0.760
Age, years	39.91	(11.84)	39.43	(11.95)	40.19	(11.76)	0.76	0.031
Experience, years	17.30	(11.99)	16.92	(11.98)	17.53	(11.99)	0.61	0.087
Earnings, € 1000	15.75	(21.31)	15.94	(22.20)	15.64	(20.76)	-0.30	0.629
Income, € 1000	31.83	(21.78)	31.91	(22.07)	31.78	(21.61)	-0.13	0.855
Reduced work capacity	0.18	(0.50)	0.47	(0.50)	0.49	(0.50)	0.02	0.110
Mental health diagnosis	0.10	(0.30)	0.10	(0.30)	0.11	(0.31)	0.01	0.294
Musculoskeletal diagnosis	0.06	(0.24)	0.06	(0.23)	0.07	(0.25)	0.01	0.302
N	4898		1825		3073			

Note: Experience is measured as years with positive labor earnings up to and including the year before randomization. Earnings and income are yearly measures, measured the year prior to randomization. Diagnoses measursd ever registered with a diagnosis up to the month before randomization.

To further assess balance, we estimated a regression model where treatment status is regressed on the baseline characteristics from Table 2, controlling for county-by-month fixed effects:

$T_i = \beta X_i + \alpha_{ij} + \varepsilon_i$, where T_i is treatment status, X_i is the vector of controls, and α_{ij} denotes the county*month fixed effects. The results from this exercise reveal that a slight imbalance – being treated was somewhat correlated with age and prior experience and earnings. A joint F-tests of the coefficients did not reject the null hypothesis of no overall imbalance. Detailed results are reported in Table A2 in the Appendix.

In addition to the administrative data, qualitative data were collected through surveys and interviews with team leaders and supervisors ("job specialists") from both treatment and control groups. Surveys were conducted at two points in time: after one year of program implementation

[2] Information about the data and instructions on how to apply for access can be found on: https://www.nav.no/no/nav-og-samfunn/kunnskap/data-og-forskning-pa-nav

and at the end of the study period. The purpose was to gather insights into provider behavior, organizational practices, and perceptions of the payment models, which could help contextualize the quantitative findings. The surveys included questions on organization, routines, quality development, content of follow-up, educational background, and perceptions of how the payment scheme affected service delivery. As part of the evaluation, group and individual interviews were conducted to explore these themes in greather depth.

5 Results

5.1 Compliance

Figure 1 shows the extent to which individuals actually participated in the programs they were randomized to. Panels (a) and (b) show that the time pattern of compliance was quite similar between the treatment and control group. Panel c) shows that around 75-80% of participants eventually participated in the program they were randomized.

Non-compliance reportedly arose from several factors, including participants securing employment, enrolling in education, becoming ill, or moving while waiting for program start-up. In some cases, participants did not show up. These reasons are not expected to differ systematically between the two groups. However, panel (c) shows that compliance was around 5 percentage points lower for individuals in the treatment group. It would be concerning if this difference indicated "cream-skimming", i.e. the providers with performance pay managing to sort out participants with low probability of success. To investigate this possibility, we examined whether the probability of participating in the assigned program varied by individual characteristics depending on which program a person was assigned to. Specifically, a variable equal to 1 if the person actually participated in the assigned program within 12 months, and 0 otherwise, was regressed on pre-treatment individual characteristics with and without interaction with an indicator for program assignment. The results from this analysis are shown in Table A3 in the appendix. Age and work experience had a statistically significant positive association with compliance for participants assigned to the control program. However, this association was weaker, or even reversed, for participants assigned to the treatment program. This is contrary to what we would expect if treatment providers systematically rejected "weaker" participants. There is, however, another pattern that is more strongly linked to actual participation across assignment groups: individuals with reduced work capacity randomized to the control program were slightly more likely to participate, while they were less likely to participate if assigned to the treatment program. This may reflect that individuals with reduced work capacity were perceived to have a lower likelihood of achieving the outcome milestones tied to at least 50% work required under performance pay contracts, an issue that was brought up in interviews.

None of these patterns affect the intention-to-treat analysis, which includes all participants as assigned. However, they suggest that there may be some degree of positive selection into the performance pay contracts, by both NAV caseworkers and providers, in a normal non-

experimental situation. We proceed with the main intention-to-treat estimates. Attrition is minimal due to the use of administrative data, which captures outcomes for all individuals regardless of program participation.

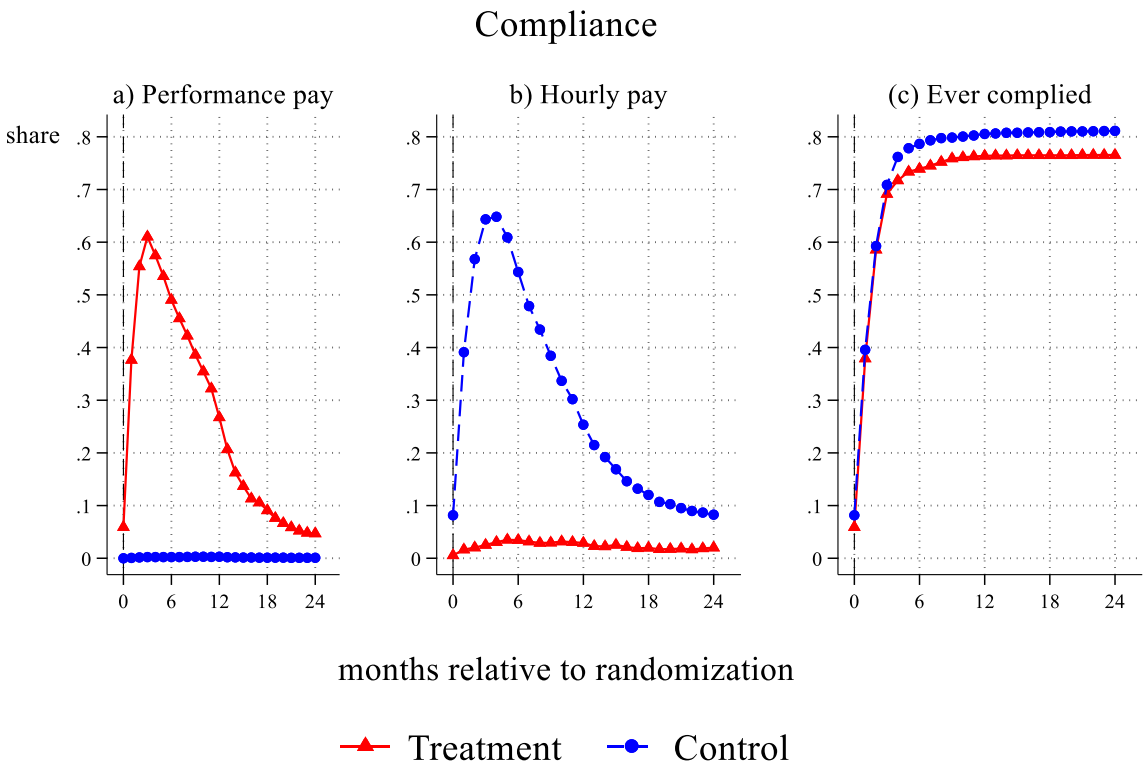


Figure 1. Actual program participation by treatment status.

5.2 Main treatment effect

The results are presented graphically for ± 24 months relative to the month of randomization and with estimates at 12 months. Treatment effects are estimated with the model:

$Y_i = \beta\, T_i + \alpha_{ij} + \gamma\, X_{i,t0} + \varepsilon_i$, where T_i is the treatment status and α_{ij} is the county-by-month fixed effects. County-by-month fixed effects are included because randomization took place within counties and the share of people drawn varied over time within counties.

The main estimation uses the post-double selection LASSO approach (Belloni, Chernozhukov, and Hansen, 2014) to optimally select control variables $X_{i,t0}$ from the complete list of controls. All potential control variables are measured before randomization. The post-double selection LASSO method identifies variables that are significantly correlated with both the treatment and the outcome. Selecting these relevant control variables can enhance the precision of the treatment effect estimates. It also helps address imbalances across the treatment and control groups that may have arisen due to chance or other factors. This approach ensures that only the most relevant control variables are included, increasing the robustness and validity of the findings. Candidate

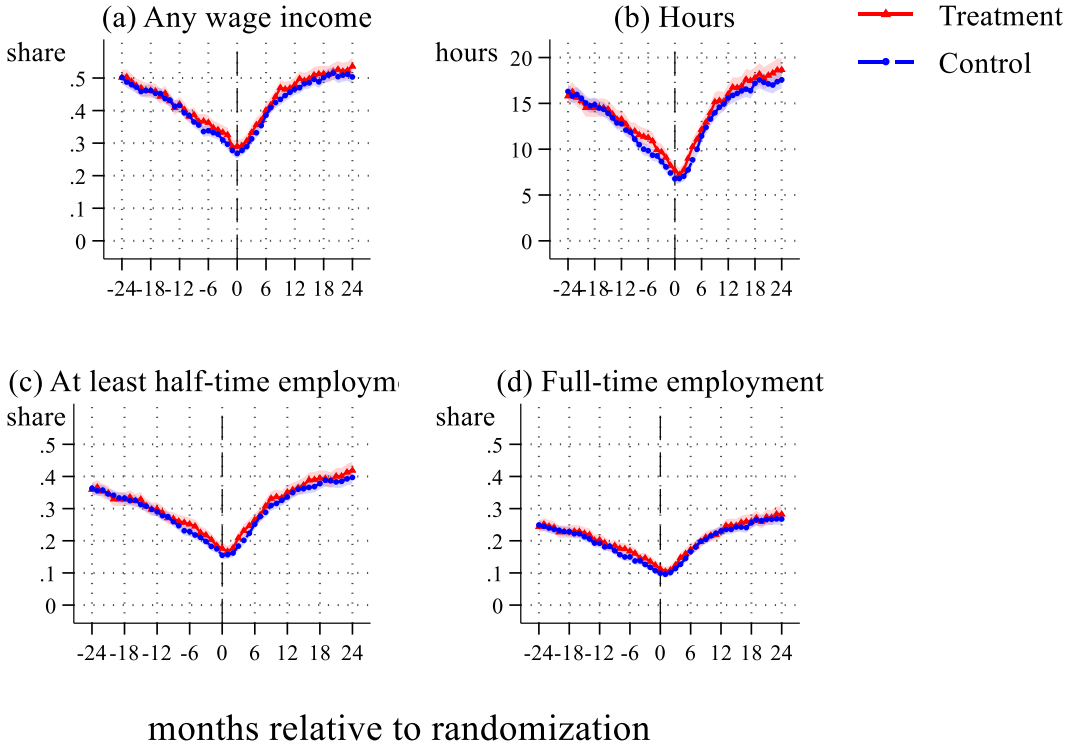
variables include binary indicators for gender and immigrant background, and continuous measures of age, experience, earnings, and income, along with all pairwise interactions between these variables. The actually selected control variables are indicated in the results table. Conventional standard errors are used, as randomization was at the individual level.

We focus on intention-to-treat effects, i.e. all individuals randomized to treatment will be included in the treatment group.

Figure 2, Panel A) shows monthly employment measures for the treatment and control groups in a time window of two years before and two years after randomization. The similarity of the levels and trends over time stands out and already indicates a close to zero effect. Table 3, Panel A, presents the estimated treatment effects on various employment outcomes at 12 months after randomization. The coefficients on the treatment indicator are small and not statistically significant across all outcomes, including earnings, employment status, hours worked, and measures of full-time employment. We can rule out effects on monthly earnings of ± €5 and employment effects of ± 1 percentage points for participants allocated to a provider with performance pay.

Figure 2, Panel B) shows our measure of job retention – the average cumulative months of employment since the month of randomization for the treatment and control. Here the similar development of the two groups is even more striking. Table 3, Panel B) shows effect estimates at 12 months, again small and not statistically different from zero.

A) Job attainment



B) Job retention

Figure 2. Job attainment among treatment and control groups. Raw means by month relative to randomization month.

Table 3. Intention to treat estimates of job attainment (Panel A) and job retention (Panel B) at 12 months after randomization.

	(1) Earnings, € 1000	(2) Any earnings	(3) Working hours	(4) At least 50% of full-time employment	(5) Full-time employment
A) Job attainment					
Treatment	0.0024	-0.0043	-0.18	0.0083	-0.091
	(0.0061)	(0.016)	(0.73)	(0.015)	(0.013)
Female	No	No	Yes	No	Yes
Age	No	Yes	Yes	Yes	No
Earnings, t-1	No	No	Yes	No	No
Income, t-1	No	No	No	No	No
Mean (control group)	0.12	0.47	15.6	0.34	0.23
N	4898	4898	4898	4898	4898
B) Job retention					
Treatment	0.050	0.11	0.12	0.16	0.046
	(0.044)	(0.14)	(0.14)	(0.12)	(0.11)
Female	No	Yes	Yes	No	Yes
Age	No	Yes	Yes	Yes	Yes
Earnings, t-1	Yes	Yes	Yes	Yes	Yes
Income, t-1	No	No	No	No	No
Mean (control group)	1.10	4.86	3.99	3.19	2.10
N	4898	4898	4898	4898	4898

Note. Treatment effects estimated using post-double selection LASSO (Belloni, Chernozhukov and Hansen, 2014). The variables available for selection include female and immigrant background (dummy variables), age, experience, earnings and income (continuous measures), as well as all interactions between these variables. The actually selected control variables are indicated with "Yes". Standard errors in parentheses.

5.3 Heterogeneous effects

We planned three different ways to estimate heterogeneous effects – subgroup analysis, the interacting characteristics with treatment, and a machine learning approach.

Jobseekers with full work capacity are likely to have higher probability for a position for at least 50% and thus be more attractive for providers with performance pay. We planned to estimate effects by NAVs "effort groups," which are categories for varying degrees of assistance needs. However, this information was not available, so we decided to use the reduced-capacity distinction as a proxy. From our experience, NAV also considers this distinction important and routinely monitors these subgroups.

Estimating effects separately by year is interesting because providers may learn to optimize services under a new pay scheme over time, so treatment effects could grow (or attenuate) as firms refine their practices and caseworkers gain experience with performance incentives. Second, the COVID-19 pandemic altered labor-market conditions and service delivery – lockdowns, remote counselling, and shifting employer demand may have disrupted placement dynamics in 2020–21 relative to 2018–19. Year-by-year estimates therefore reveal both any learning-curve in incentive uptake and the extent to which pandemic shocks modified the scheme's effectiveness.

Figures 3 and 4 show graphically that there are no signs of differential effects by work capacity or in the different years of the trial.

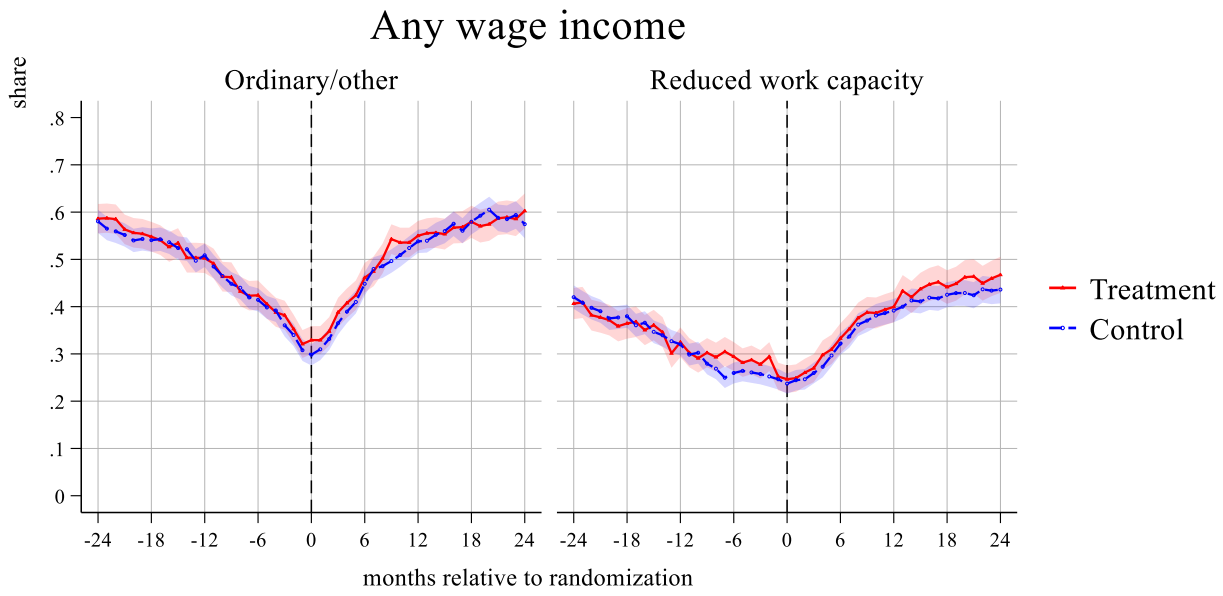


Figure 3. Job attainment by work capacity and treatment status.

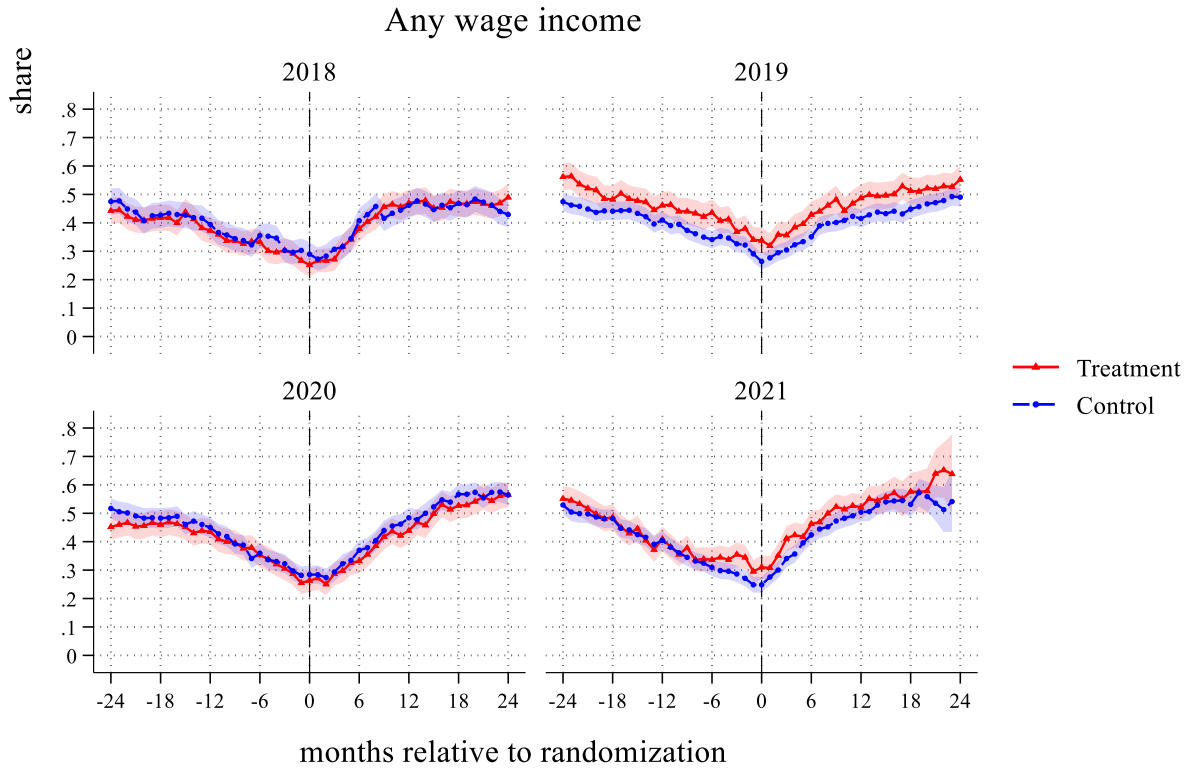


Figure 4. Job attainment by year of randomization and treatment status.

We also investigated heterogeneous treatment effects with the traditional interaction approach, examining whether the effect differed across the characteristics gender, immigration status, age, experience, prior earnings, and prior income. The results indicated no significant differences along these dimensions.

As there are some drawbacks to the interaction approach, we also employed the "causal forest" method to automate the search for heterogeneous treatment effects (Wager and Athey, 2017; Athey, Tibshirani, and Wager, 2019). This method involves dividing the sample into training and test data and using the training data to create trees that split covariates in a way that maximizes the difference in treatment effects between child nodes. By averaging over 5000 of these trees, a forest was created, yielding estimated Conditional Average Treatment Effects (CATE) for each individual. All individuals were then categorized into low and high impact groups based on whether they were below or above the median value. Figure A2 in the appendix shows average actual outcomes in the test data by treatment status for these two groups. The difference between the treatment and control group is almost identical in both groups, indicating no evidence of heterogeneous treatment effects.

5.4 Cost-benefit analysis

On the benefit side, there was no discernible difference between the two schemes in terms of job attainment and retention rates; these outcomes were statistically similar. We also assumed that other potential benefits were equal between the two programs.

As described in section 3.3, both the performance-pay and hourly-fee schemes were managed under NAV's existing procurement and oversight framework. Both schemes therefore involved administrative costs. Some interviewees reported higher administrative costs for performance pay, however, there was agreement that these were attributable to the trial itself. Under normal operations, NAV's county offices expected similar contracting, monitoring, and caseworker time commitments across both schemes. We therefore consider the administrative costs to be the same across the two schemes.

The actual prices submitted in the bids for the performance pay contracts were confidential. Similarly, we did not have access to prices in the regular contracts. While contract-level data were unavailable, NAV provided aggregate accounting figures on payments to providers in the period 2019-2022. These numbers show that providers with performance pay contracts were paid approximately €467 per participant per month, compared to approximately €640 per participant per month for providers in the control group. Although subject to considerable uncertainty, these estimates clearly suggest that the performance pay scheme was less costly, with an average saving of around €1600 for the average participant.

There is a question of whether the trial payment bids would be the same in ordinary operations, and whether those prices accurately reflect the underlying resources use. When providers bid to deliver under the performance pay contract, they faced considerable uncertainty – in setting their prices, they had to estimate what share of participants they could place into employment and thus bill for. The interviews suggested that some providers were too optimistic and would have needed to raise prices in a future tender. Others achieved the returns they expected, while a few intentionally bid low to "enter the market." Overall, there is insufficient information to determine whether the performance pay contract prices accurately reflected resource use.

In conclusion, the estimated cost saving of €1600 per participant is substantial and lends support to incentive-based contracting. However, given the uncertainty in the underlying cost estimates, one should be cautious in drawing strong conclusions about relative cost-effectiveness.

5.5 Provider behavior and perceptions

The survey results showed few systematic differences between the two groups in terms of organization, routines, quality development, and the content of participant follow-up. Both groups reported similar approaches to service delivery, suggesting that the payment scheme did not lead to substantial changes in these areas.

However, among supervisors in the treatment group, several notable differences emerged. A higher share of these supervisors had higher college or university education compared to those in the control group. Treatment group supervisors reported using "work training" less frequently before placing participants into actual employment, instead focusing more on direct job placements. They also reported that that they conducted more thorough assessments and job matching, tailoring their support more closely to individual participant needs and that they felt more flexible in how they allocated time for each participant, allowing them to prioritize resources based on participants' readiness and potential for employment.

When asked specifically about the payment model, treatment group supervisors said that the performance-based payment scheme improved quality and staff competence, lead to longer and more thorough follow-up and inspired "smarter work". These views were reinforced in group and individual interviews, where supervisors elaborated on how the payment scheme influenced their practices.

Providers were free to pass the financial incentives in the funding scheme on to their staff through individual or collective bonus schemes or other rewards for counsellors. The interviews showed that three of the treatment providers had bonus arrangements during the trial period. In one case, initially only the manager; however, the manager chose to discontinue it, explaining that it distorted work focus. After that, the provider had no bonus scheme for staff. A second treatment provider began with an individual bonus for counsellors but later switched to a collective bonus, stating that they "wanted to think more holistically." The third provider implemented a collective bonus from the outset, to be shared among all employees, justifying this by noting that counsellors receive very different participant profiles, some ready to enter employment immediately, others requiring years of follow-up. Whether or not they had a bonus scheme, almost all staff were well aware of the funding incentives. In 2021, 17 of 18 job specialists reported knowing the details of the program's financing model; 13 of those 17 said it was discussed in meetings, and 6 said it was also discussed informally among counsellors outside of meetings.

Despite these perceived differences in approach and attitudes, these changes did not translate into measurable differences in participant employment outcomes, as evidenced by our quantitative analysis.

5.6 Program participation dynamics

An analysis of program participation revealed a distinct pattern in the timing of exits between the treatment and control groups. Among participants who started their assigned program, those in the performance pay (treatment) group exited the program slightly quicker in the first six months compared to those in the hourly pay (control) group. After six months, however, a larger share of participants in the treatment group remained in the program compared to the control group.

Figure 5 shows the cumulative exit rates from the program over time for both groups. The faster early exit rate in the treatment group may suggest that providers operating under performance-

based contracts were incentivized to prioritize participants who could achieve employment milestones quickly. This behavior is consistent with the concept of "cream-skimming," where providers focus their efforts on participants who are more likely to generate results-based payments due to their higher employability (Koning and Heinrich, 2013; Carter and Whitworth, 2015).

Conversely, the larger share of participants remaining in the program after six months in the treatment group could indicate "parking" behavior. Providers might allocate fewer resources to participants who are less likely to meet the employment milestones within the contract period, effectively keeping them in the program without substantial progress (Koning and Heinrich, 2013). Alternatively, it may reflect providers investing in longer-term support for participants requiring more intensive assistance, as suggested by the supervisors in the treatment group who reported providing longer and more thorough follow-up (see Section 5.5).

These patterns suggest that performance-based incentives may have led providers to tailor their strategies to participants' perceived employability. Providers might concentrate resources on job-ready participants to secure quick outcome payments while deprioritizing those needing more support. Such strategic behavior could result in unequal service provision, potentially disadvantaging participants with greater barriers to employment. However, it is hard to know whether this was actually the case, and the pattern is also consistent with other explanations.

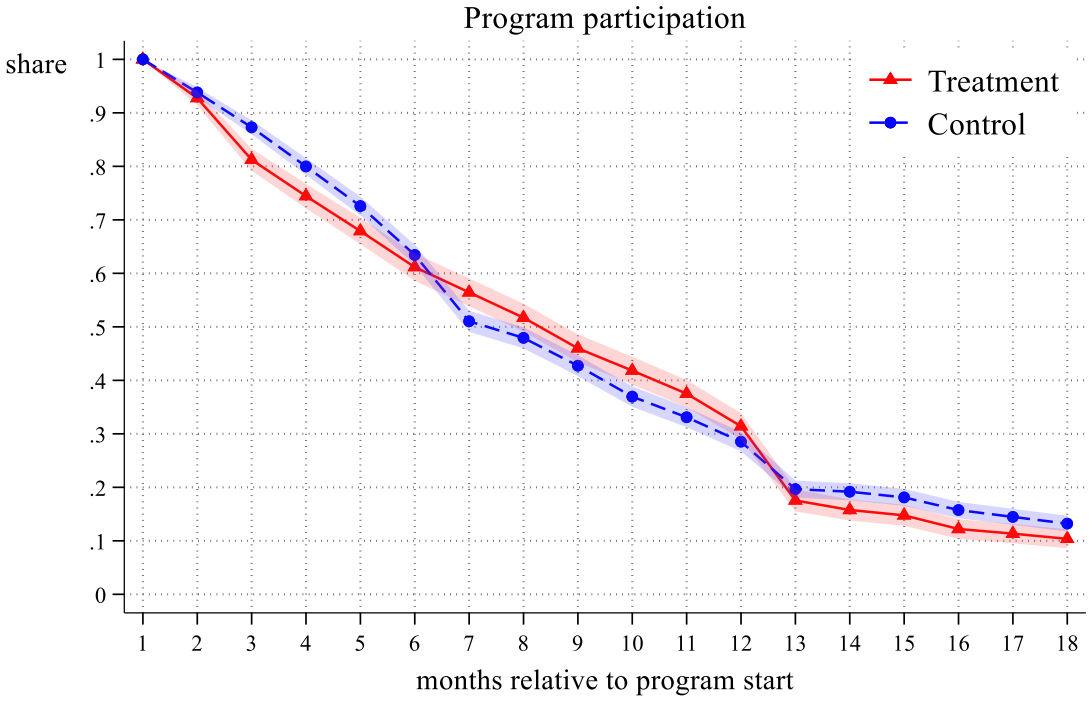


Figure 5. Share still in the program, conditional on starting.

6 Conclusion

The field experiment found no significant effects of performance-based pay for private employment service providers on hours worked, earnings, or employment rates. An accompanying cost-benefit analysis, though subject to uncertainty, indicated positive results through cost savings.

Several alternative explanations could account for these null findings. One possibility is that the standard hourly pay program was already highly effective, leaving little room for improvement through altered incentive structures. Providers under both payment schemes may have delivered similar services due to professional norms, regulatory requirements, or intrinsic motivations, thereby diluting the impact of financial incentives. It is unlikely that the performance-based incentives were too weak to elicit changes in provider motivation.

The absence of heterogeneous effects suggests that performance-based pay did not differentially benefit or disadvantage any particular group. External factors such as the COVID-19 pandemic do not appear to explain the findings, as the results are consistent across pre-pandemic and post-pandemic periods.

Qualitative data from surveys and interviews provide context for our findings. Supervisors working under the performance pay scheme perceived that the incentives led to improvements in service delivery, including higher quality and competence, more thorough assessments and job matching, and increased flexibility in supporting participants. They also felt inspired to work smarter and believed that the payment scheme positively impacted their practices. However, these perceived enhancements did not manifest in improved employment outcomes for participants. This suggests that while provider behavior and attitudes may have been influenced by the payment scheme, other factors, such as structural constraints, participant characteristics, or broader labor market conditions, may have limited the impact on actual employment results. The lack of significant differences in organization, routines, and content of follow-up between the two groups indicates that providers under both payment schemes were delivering services in similar ways.

From a policy perspective, these findings indicate that shifting to performance-based contracts for private employment service providers may not yield the desired improvements in employment outcomes, although the results do suggest some degree of cost savings. Policy makers should nevertheless also consider improving other aspects of program design, provider selection, or participant support.

It is important to acknowledge the study's limitations. The non-random assignment of providers to payment schemes could introduce unobserved differences between providers. The specific design of the performance pay contract, including the size and structure of incentives, may also influence its effectiveness. Future research could explore alternative incentive structures, larger or more immediate payments, or complementary interventions to enhance provider responsiveness.

In conclusion, while performance-based pay for private providers did not improve employment outcomes in this context, the insights gained contribute valuable knowledge to the discussion on

optimal contracting in public service outsourcing. Further is needed on strategies that effectively align provider incentives with policy objectives to enhance the efficacy of labor market programs.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work the author used ChatGPT in order to improve language and readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Andersson, Fredrik, Henrik Jordahl, Jens Josephson (2019) Outsourcing Public Services: Contractibility, Cost, and Quality, *CESifo Economic Studies*, Volume 65, Issue 4, December 2019, Pages 349–372, https://doi.org/10.1093/cesifo/ifz009

Athey, Susan, Stefan Wager (2021) Policy learning with observational data. *Econometrica*, 89(1), 133-161. https://doi.org/10.3982/ECTA15732

Athey, S., Tibshirani, J., Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2), 1148–1178. https://doi.org/10.1214/18-AOS1709

Bennmarker, Helge, Erik Grönqvist, Björn Öckert (2013) Effects of contracting out employment services: Evidence from a randomized experiment, Journal of Public Economics, Volume 98, Pages 68-84, https://doi.org/10.1016/j.jpubeco.2012.12.002.

Blank, R. M. (2000). When can public policy makers rely on private markets? The effective provision of social services. *The Economic Journal*, *110*(462), 34-49. https://doi.org/10.1111/1468-0297.00519

Bredgaard, Thomas, and Flemming Larsen (2008) Quasi-markets in employment policy: do they deliver on promises?, *Social Policy and Society* 7/03. https://doi.org/10.1017/S1474746408004314

Bruttel, O. (2005) *Contracting-out and governance mechanisms in the public employment service* (No. SP I 2005-109). WZB Discussion paper.

Carter, Eleanor, Adam Whithworth (2015) Creaming and Parking in Quasi-Marketised Welfare-to-Work Schemes: Designed Out of or Designed In to the UK Work Programme? *Journal of Social Policy* 44.02, 2015. https://doi.org/10.1017/s0047279414000841

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

Crépon B. (2018) Private providers of labor market services: a review of the evidence, in Bergström, A and L Calmfors (eds.), *Framtidens arbetsförmedling*, Fores, Stockholm.

Domberger, S., Jensen, P. (1997) Contracting out by the public sector: theory, evidence, prospects. *Oxford review of economic policy*, *13*(4), 67-78.

Econ Pöyry, Proba samfunnsanalyse (2010) Alternative finansieringsformer for Arbeid med bistand. *Econ-rapport R-2010-068*.

Egebark, J., Laun, L., Liljeberg, L., Rödin, M., Söderström, M., Videnord, E., Vikström, J. (2024). *En effektutvärdering av arbetsförmedling med fristående leverantörer*. IFAU-Institute for Evaluation of Labour Market and Education Policy, Rapport 2024: 17.

Hart, O., Shleifer, A., Vishny, R. W. (1997). The proper scope of government: theory and an application to prisons. *The Quarterly Journal of Economics, 112*(4), 1127-1161. https://doi.org/10.1162/003355300555448

Homrighausen, P. (2014). *Differential pricing and private provider performance,* IAB-Discussion Paper No. 25/2014.

Koning, Pierre, Carolyn J. Heinrich (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. https://doi.org/10.1002/pam.21695

Koning, P., Van De Meerendonk, A. (2014). The impact of scoring weights on price and quality outcomes: An application to the procurement of Welfare-to-Work contracts. *European Economic Review*, *71*, 1-14. https://doi.org/10.1016/j.euroecorev.2014.06.017

Langenbucher, K., M. Vodopivec (2022) Paying for results: Contracting out employment services through outcome-based payment schemes in OECD countries, *OECD Social, Employment and Migration Working Papers*, No. 267, OECD Publishing, Paris. https://doi.org/10.1787/c6392a59-en.

Laun, L., Peter Skogman Thoursie (2014) Does privatisation of vocational rehabilitation improve labour market opportunities? Evidence from a field experiment in Sweden, *Journal of Health Economics*, Volume 34, 2014, Pages 59-72, ISSN 0167-6296. https://doi.org/10.1016/j.jhealeco.2013.12.002.

Lunder, Trond Erik, Knut Løyland, Geir Møller, Joar Sannes (2005) Evaluering av nye finansieringsformer i Aetat. *Rapport 225. Bø: Telemarksforsking*.

Meld. St. 33 (2015-2016) *NAV i en ny tid – for arbeid og aktivitet*. Stortingsmelding 33 (2015-2016).

NOU (2012) *Arbeidsrettede tiltak*. Report delivered to the Ministry of Labour, February 2012, NOU: 2012:6.

Proba samfunnsanalyse (2014) Forsøk med resultatbasert finansiering av formidlingsbistand - Kartlegging av oppstartfasen. *Rapport* 2014-12.

Proba samfunnsanalyse (2017) Evaluering av offentlig anskaffelse – Delrapport 1. *Rapport 2017-01*.

Proba samfunnsanalyse (2023) Evaluering av Refino – Forsøk med resultatbasert finansiering av oppfølgingstiltak1. *Rapport 2023-15.*

Rehwald, Kai, Michael Rosholm, Michael Svarer (2017) Do public or private providers of employment services matter for employment? Evidence from a randomized experiment, *Labour Economics*, Volume 45, 2017, Pages 169-187, https://doi.org/10.1016/j.labeco.2016.11.005

Stephan, G. (2024). Public or private job placement services—Are private ones more effective?. *IZA World of Labor 2024: 285* http://dx.doi.org/10.15185/izawol.285.v2

Appendix

County	Region	Provider with performance pay	Provider with hourly pay	Change of provider with hourly pay
Akershus	**Ullensaker and Eidsvoll**	Din utvikling (FP) (Runner-up: Fretex (NP), Arbeidslivsressurs (FP), OTTO Ressurs (FP))	Sens Utvikling (FP)	No change
Hordaland	**Bergen and Nord-Hordland**	Din utvikling (FP) (Runner-up: Arbeidslivsressurs (FP), Fretex (NP), A2G)	Fretex (NP)	Din utvikling (FP) from August 2020
Oslo	**Vest**	AS3 Employment (FP) (Runner-up: Arbeidslivsressurs (FP), Din utvikling (FP), Fretex (NP))	Din utvikling (FP)	No change
Telemark	**Grenland and Vestmar**	Arbeidslivsressurs (FP) (Runner-up: GREP arbeid (FP), JobLearn (FP), OTTO (FP))	Region 1: KEOPS (M) Region 2: AS3 employment (FP)	Region 1 and 2: Sonans (FP) from July 2020
Østfold	**Sarpsborg**	Fretex (NP) (Runner-up: Bedriftsakademiet (FP), Arbeidslivsressurs (FP), Din utvikling (FP)	Din utvikling (FP)	Sens utvikling (FP) from July 2019

Table A1. Program providers. FP: For profit (commercial), NP: Non-profit, M : Municipal business enterprise.

Table A2. Balance

	Treatment									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(7)
Female	0.021									0.022*
	(0.013)									(0.013)
Immigrant		-0.0036								-0.0054
		(0.015)								(0.020)
Age			0.0011**							0.0012
			(0.00056)							(0.0012)
Experience				0.00099*						-0.0004
				(0.00055)						(0.0013)
Earnings, € 1000					0.00068**					0.00043
					(0.00031)					(0.0005)
Income, € 1000						0.00062**				0.0015
						(0.00030)				(0.0005)
Reduced work cap							-0.012			-0.016
							(0.013)			(0.015)
Mental, health								0.0031		0.016
								(0.021)		(0.023)
Musculo-skeletal									0.031	0.032
									(0.027)	(0.028)
Control mean	0.54	0.29	40.19	17.53	156	317	0.49	0.11	0.07	
N	4898	4898	4898	4898	4898	4898	4898	4898	4898	4898
R-squared	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
County*month	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Joint F-test										0.1756

Note: Results from a regression of treatment (1 if person is randomized to the treatment program) on characteristics. Variables measured the year before randomization. Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001

Table A3. Compliance

	Compliance						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	0.0262						0.0116
	(0.0135)						(0.0147)
X Treatment	-0.0487**						-0.00591
	(0.0172)						(0.0238)
Immigrant		0.0166					-0.00598
		(0.0159)					(0.0227)
X Treatment		-0.0479*					0.0197
		(0.0234)					(0.0363)
Age			0.00128*				0.00136
			(0.000530)				(0.00114)
X Treatment			-0.00109***				-0.00182
			(0.000319)				(0.00131)
Reduced work cap.				0.0253			0.0181
				(0.0139)			(0.0155)
X Treatment				-0.0686***			-0.0478
				(0.0184)			(0.0245)
Experience					0.00219*		-0.000880
					(0.000559)		(0.00135)
X Treatment					-0.00144*		0.00306
					(0.000620)		(0.00200)
Earnings						0.00000547	0.00000306
						(0.00000341)	(0.00000381)
X Treatment						-0.0000107*	0.00000612
						(0.00000470)	(0.0000590)
N	4898	4898	4898	4898	4898	4898	4898
R-squared	0.071	0.070	0. 072	0. 072	0. 071	0. 070	0. 074
County*month	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: Results from a regression of compliance (1 if person is registered in the assigned program within 12 months of randomization) on characteristics. Variables measured the year before randomization. Earnings measured in €1000 Standard errors in parentheses. * p<0.05, ** p<0.01, *** p<0.001

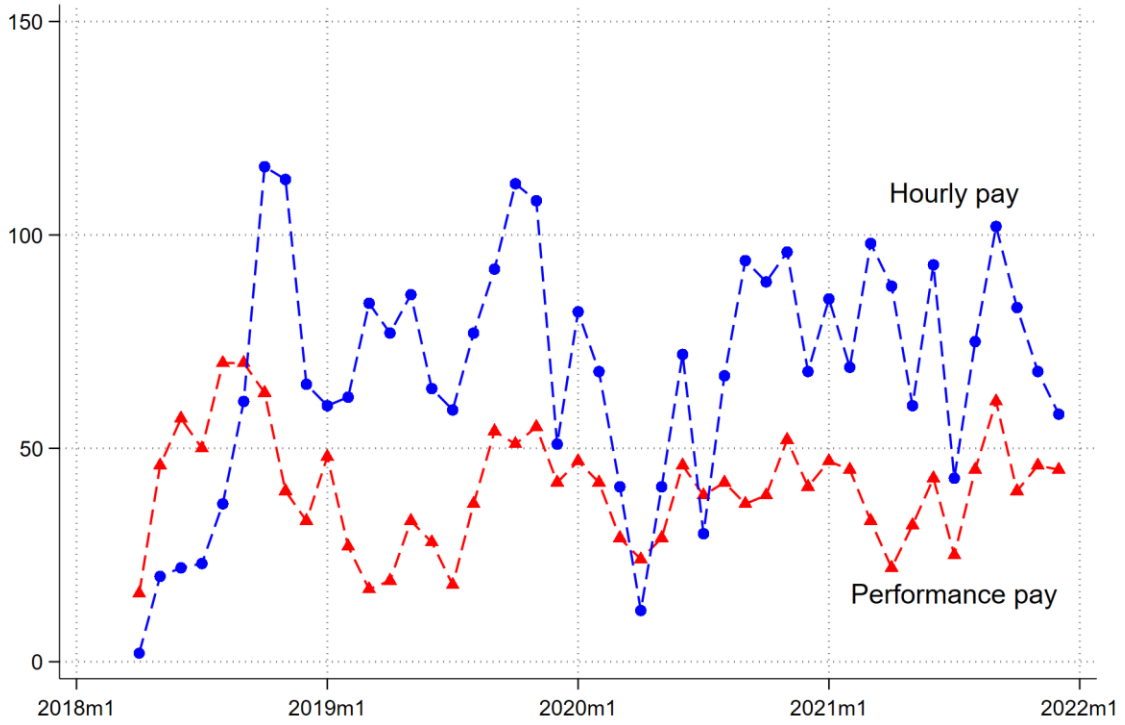


Figure A1. Randomized participants. 3073 to hourly pay (blue), 1825 to performance pay (red).

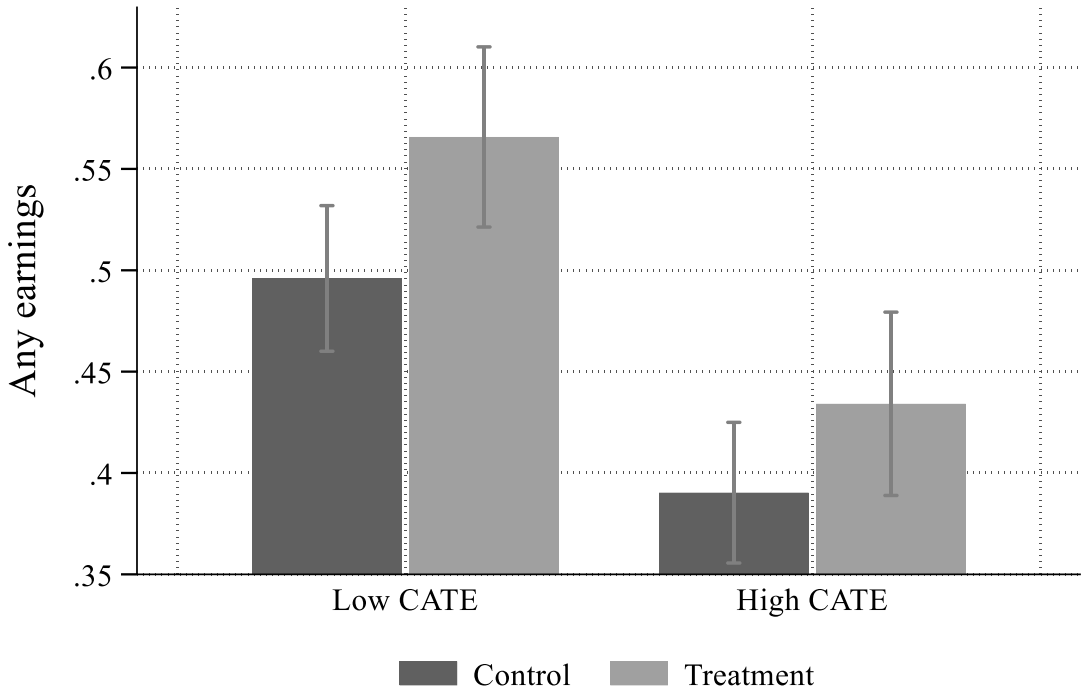


Figure A2. Any earnings by treatment status and predicted CATE group (test sample).