Cash Transfers, Health, and the Transition to Adulthood

Ossi Tahvonen*[1]

[1] *University of Helsinki & Helsinki Graduate School of Economics*

November 5, 2026

For the most recent version, click here

Abstract

This paper studies how a health-based cash transfer conditional on schooling affects adolescents' long-run labor market outcomes. To do so, I use administrative data and quasi-random assignment of applicants to examiners who differ in leniency. Among marginal applicants—those whose approval depends on the assigned examiner—approval for the benefit reduces the probability of working or studying after payments end by 12 percentage points. I find that the program did not increase the marginal applicants' educational attainment but weakened their early attachment to the labor market. To assess the effects on the wider population of applicants, I use comparisons of observably similar approved and rejected applicants. This analysis suggests that the program improves later employment among adolescents with the poorest labor market prospects, shifting them from disability benefits into education during the payment period. My findings demonstrate that the effects of education-based cash transfers in high-income countries can depend critically on the target population.

Keywords: disability benefits, conditional cash transfers, labor supply, mental health

JEL: I38, J13, J14, J22

*I am thankful for the funding this project has received from the Yrjö Jahnsson Foundation, Liikesivistysrahasto, Stone Center, and Kuntoutustukisäätiö Peurunka. I thank Torsten Santavirta, Roope Uusitalo, and Pedro Carneiro for their generous guidance. This work was also substantially improved by comments from Mikko Silliman, Matti Sarvimäki, Dan-Olof Rooth, Jukka Pirttilä, Jerry Montonen, Miika Päällysaho, Marc Riudavets, Ellen Sahlström, Jussipekka Salo, Matti Mitrunen, and Saara Haikonen. I also thank seminar/conference audiences at Helsinki GSE, Stockholm University, the Frisch Centre workshop on Marginalized Young Adults, NEWER at the Umeå University, HEEW 2026 and HEEPs for comments.

1 Introduction

Across high-income countries, mental and physical health impairments affect a substantial share of adolescents. In the EU, 8.1% of 16–29-year-olds report health-related limitations in usual activities (Eurostat, 2026). These adolescents face barriers to entering the labor market, and many drop out of education early, instead ending up receiving passive disability benefits that can weaken incentives to work (Maestas et al., 2013; Deshpande, 2016a; Autor et al., 2019; Levere, 2021). This creates a challenge for policy: how to provide income support for this disadvantaged group while encouraging human capital accumulation at this formative age.

One response is to provide cash transfers conditional on continued education. Such transfers are a popular policy tool since they can alleviate resource constraints while incentivizing education which might prove useful in adulthood. Yet subsidizing participation in education can have heterogeneous effects, some of them unintended. If the transfer is targeted towards a group for which the schooling condition is not binding, the additional income could reduce work alongside schooling and displace valuable labor market connections without changing school attendance. Thus, the overall effects of this type of policy on long-run outcomes could crucially depend on whom they target.

Finland offers an excellent setting to study these potentially opposing effects. As an alternative to passive disability benefits, adolescents with health impairments can receive a cash transfer conditional on continued education or another rehabilitative activity.[1] The policy is a previously unstudied combination of generous ($\sim$ 700 euros/month) income support with incentives to invest in human capital, aimed at improving labor-market outcomes in the long run. It targets adolescents (ages 16–20) with a documented health problem that threatens their education or employment prospects. The most common disabilities among the recipients are mental health and neurodevelopmental problems, ranging from intellectual disability and autism to depression, and thus covering applicants with varying labor market prospects. Some applicants would otherwise start receiving disability benefits, while others would stay in school even without the transfer.

In this paper, I study how this education-conditioned cash transfer affects applicants' education, labor-market outcomes, and health into their thirties. I link comprehensive administrative data on labor-market outcomes and healthcare use to novel records on applications, diagnoses, and examiners for around 30,000 applications between 2004 and 2018. This data allows me to conduct the first evaluation of causal effects of a program that replaces disability benefits with a conditional cash transfer for adolescents, with a follow-up until their thirties.

My main identification strategy uses the fact that the applications for the benefit are allocated to examiners quasi-randomly from a running queue in an Instrumental Variables (IV) design. Examiner leniency varies substantially: a one-standard-deviation increase in examiner leniency increases the probability of approval by 4.9 percentage points. The instrument is uncorrelated with the applicant's background characteristics and passes a variety of other validity checks. The effects estimated are local to the group of marginal applicants: those approved by some examiners but rejected by others. This group has

[1] In practice, around 85% of recipients attend secondary education.

better labor-market prospects than the average applicant to the benefit.

I find economically meaningful adverse labor-market effects among marginal applicants. Contrary to the policymakers' aim, benefit approval reduces employment by 11 percentage points and increases the probability of being in not-in-education-employment-or-training (NEET) by 12 percentage points after the benefit ends. These estimates correspond to 28% and 21%, respectively, of the corresponding means without approval. Approval for the benefit does not increase the probability of continuing studies beyond secondary education, and I find no statistically detectable effects on serious adverse health events, criminal convictions, or financial distress. For marginal applicants, the IV estimates therefore provide no evidence that approval improves long-run labor-market attachment or produces offsetting gains in the other measured domains.

Most marginal applicants would have continued in education even without the transfer. I show that approval does not materially change their time in education during the 3–4 years when the benefit is paid. However, it reduces work: cumulative earnings over the benefit period fall by 36%, likely as a result of less part-time or summer work alongside schooling. For these applicants, the transfer thus appears to displace early work rather than induce additional education. The loss of early work experience and job connections may help explain their lower employment after the benefit ends.

The IV estimates concern marginal applicants with relatively good labor market prospects. In the second part of the paper, I complement the IV analysis by investigating the effects across the applicant population, particularly among adolescents with poor prospects who are at higher risk of entering disability benefits. I combine detailed application records with pretreatment information on applicants' backgrounds and use double/debiased machine learning (DML) to adjust for selection into approval (Chernozhukov et al., 2018). To assess the credibility of the selection-on-observables approach, I benchmark its estimates against the IV estimates in the complier-weighted sample. The selection-on-observables estimates closely match the IV estimates, remain stable across alternative covariate sets, and pass standard balance and placebo tests.

I find that effects vary sharply with predicted labor-market prospects. Among applicants with the lowest prospects, approval raises employment by 4.7 percentage points from a mean of 16% and lowers NEET status. This group mainly has disabilities like intellectual disability and chromosomal disorders, which are clearly diagnosable and often lead to early receipt of disability benefits. During the benefit period, their time on disability benefits falls by more than two years, and their time in education rises, which may help explain their higher employment after the benefit ends.

What explains these heterogeneous effects? I provide a simple human-capital framework in which education and work experience or work attachment in adolescence raise adult employment. I estimate the returns to both time spent in education and early work attachment from the data, then combine these returns with the estimated effects of benefit approval on both inputs to analyze the role these channels have in explaining the effects on adult labor market outcomes. This simple model can reproduce both the sign and order of magnitude of the estimates. I find little evidence for alternative channels involving substance use (e.g. Evans and Moore, 2011; Chorniy et al., 2025) or increased information causing applications to related health-based benefits (Dahl et al., 2014; Dahl and Gielen,

2021).

I also examine how benefit approval affects other family members. I find only small reductions in parental earnings, suggesting limited parental labor-supply responses to the additional income. In contrast, approval increases the number of applications and the probability of benefit receipt among siblings, suggesting that information about the program may spread within families.

To evaluate the welfare effects of the policy, I calculate the Marginal Value of Public Funds (MVPF) (Hendren and Sprung-Keyser, 2020) for the policy, inclusive of all benefits, healthcare expenses, and costs of crime. The MVPF for the marginal applicants is 0.8, indicating that the fiscal costs exceed the recipients' willingness to pay for the benefit. In contrast, providing the benefit to the population with low labor-market prospects appears to pay for itself, although there is considerable uncertainty around the point estimates.

Related literature. The key contribution of this paper is to show that a schooling-conditioned cash transfer for adolescents with health problems does not necessarily avoid the adverse effects of the traditional, more passive disability transfers. Prior to this paper, there is scarce evidence of causal effects of programs that combine cash support with human capital investment for this large and vulnerable group. I further show that the effects differ based on what applicants would otherwise have done. My results add to three literatures.

First, I add to the literature on disability benefits for adolescents by showing that these benefits can lower employment even when they are temporary and promote staying in school. Previous work has studied the effects of Supplemental Security Income (SSI), a non-temporary program in the US (Deshpande, 2016a; Deshpande and Mueller-Smith, 2022; Levere, 2021)[2]. In line with theoretical models that abstract from human capital accumulation (Diamond and Sheshinski, 1995), this evidence shows that disability benefits without incentives for human capital investment can have negative effects on labor supply but also considerable insurance value against earnings losses caused by disability. Unlike this previous research, I study a temporary non-means-tested program that also incentivizes investment in human capital, all features that leave less room for the negative labor-market consequences and a possibility for positive ones. [3] Yet these features do not resolve the trade-offs, as I observe that while helping some, this kind of benefit can still inhibit labor-market success.

Second, my results shed new light on the effects of schooling-conditioned cash transfers, helping explain how these policies can sometimes raise enrollment without raising later earnings. Prior studies, mostly of transfers targeted on family income, find that financial support tied to education increases participation or enrollment (Dearden et al., 2009; Dynarski, 2003; Denning et al., 2019; Barr et al., 2021; Britton et al., 2025), yet may not increase or even reduce subsequent earnings (Britton et al., 2025; Barr et al., 2021). My results highlight an important unintended consequence of these transfers: by reducing

[2]There is also a large literature on the effects of (mostly permanent) disability benefits for adults (Bound, 1991; Maestas et al., 2013; French and Song, 2014; Kostøl and Mogstad, 2015; Borghans et al., 2014; Autor et al., 2019; Black et al., 2024; Huynh and Chan, 2024; Haller et al., 2024; Silver and Zhang, 2022; Gelber et al., 2023). Compared to the program studied here, the target populations (adults) and incentives imposed (permanent transfers) are very different in these programs.

[3]From a social welfare perspective, the transfer is intended to solve the problem of privately optimal but publicly too low levels of investment in human capital in the presence of disability benefits for adults.

consumption-driven labor supply during education (see also Humlum and Vejlin (2013) who also find no improvements in education but a decline in short-run work), they can displace early work connections, weakening later labor-market attachment. [4]

Third, I add to the literature on active labor-market programs (ALMPs) (Card et al., 2018; Kluve et al., 2019) by showing that education-based support at the school-to-work transition improves later employment among applicants with poor labor-market prospects, for whom it mainly replaces disability benefits, and reduces it among marginal applicants, for whom it displaces early work. This is a different margin than in previous studies comparing different types of education and employment-oriented support for individuals with reduced work capacity, finding benefits from quicker employment placement but mixed longer-run advantages across approaches (Borgan Reiling et al., 2024; Markussen and Røed, 2014; Fogelgren et al., 2023)[5]. This heterogeneity is consistent with evidence that activation can offset the negative employment effects of cash support (Filippucci, 2025), and that formal education can yield substantial returns when disability would otherwise push workers out of employment (Humlum et al., 2025). My mechanism evidence also connects to research on labor-market entry conditions (Raaum and Røed, 2006; Nordström Skans, 2011; Oreopoulos et al., 2012) and on early work experience and job connections acquired during school (Le Barbanchon et al., 2020; Hensvik et al., 2023; Bruhn et al., 2025).

Beyond employment and earnings, I find no evidence of the reductions in crime (Deshpande and Mueller-Smith, 2022) or financial distress (Deshpande et al., 2021), or of large effects on health (previous evidence varies; see (Black et al., 2024; Huynh and Chan, 2024; Silver and Zhang, 2022; Gelber et al., 2023)).

Overall, this paper finds evidence that the effects of a conditional cash transfer can be heterogeneous with respect to the counterfactual activity. Importantly, this heterogeneity is based on variables observable also to the policymakers, making this more actionable. This suggests that when designing disability, education, and rehabilitation benefits, close attention should be paid to the target population and their incentives, not only the average effects.

2 Institutional background

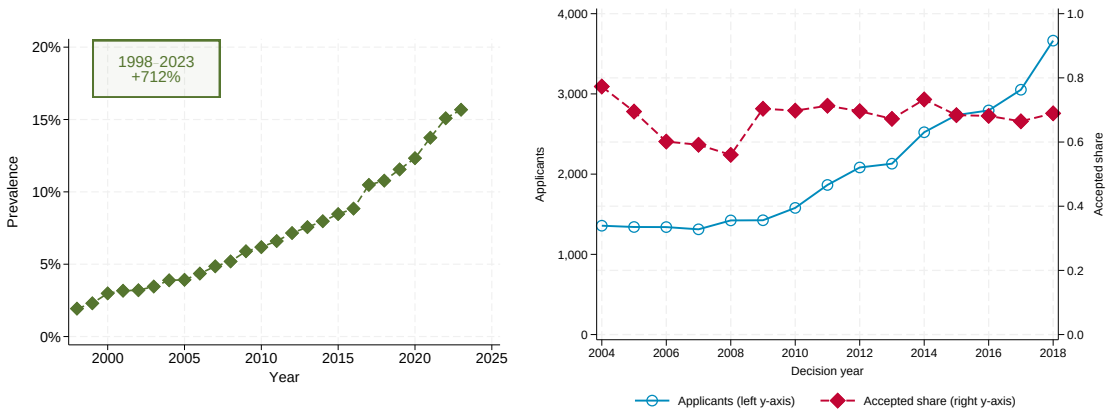
2.1 Youth Rehabilitation Allowance

The Youth Rehabilitation Allowance (YRA) is a cash benefit for 16–20-year-olds whose disability threatens their ability to study or work. It is paid and administered by KELA, the social insurance institution of Finland at the national level.

Policymakers introduced YRA as a temporary alternative to Disability Insurance (DI), which can lead to a permanent exit from the labor market. The aim was to improve recipients' long-run outcomes by making the temporary benefit conditional on participation in education or rehabilitation. If most applicants would otherwise receive DI, YRA adds

[4] In contrast, education-linked transfers and subsidies in developing countries have improved adult outcomes (Parker and Vogl, 2023; Duflo et al., 2024). In these settings, the counterfactual is more often leaving school or not enrolling, with participation in family production (Attanasio et al., 2012).

[5] Relatedly, Reiterä (2021) find that vocational-rehabilitation approval increases subsequent benefit receipt among young applicants. The paper instruments approval with examiner leniency which inspired the identification strategy in this paper.



(a) Share of 16 year olds with mental healthcare contact, years 1998-2020

(b) Number of applicants and share of accepted applicants, years 2004–2018.

Figure 1: Mental healthcare contact and YRA applications. Author's calculations.

little to public spending in the short run. It could even pay for itself if recipients claim fewer benefits as adults. Additional motivations for the benefit include the possibility that disabled individuals may face additional costs for rehabilitative treatments.

YRA may affect recipients through more than the cash itself. The program can change their incentives to attend school or rehabilitation and may lead to closer monitoring of school attendance. The application process can also attach a medical label, increase applicants' knowledge of the benefit system, and put them in contact with school and social service workers. In the rest of the paper, I refer to the treatment as a cash transfer or benefit, but I do not attempt to separate the effect of cash from these other channels.

Over time, both annual applications and the stock of YRA recipients have increased considerably (Fig. 1). The stock of YRA recipients has overtaken the pre-YRA stock of DI recipients (while DI recipiency has also continued to increase; see Fig. A.13b). Most of this growth comes from applications based on mental and behavioural disorders (ICD-10 chapter F, including intellectual disabilities and developmental disorders), which now account for approximately 89 % of applicants and 91 % of accepted applicants. This pattern mirrors a broader rise in mental health diagnoses among adolescents. Finnish birth cohorts contain about 60,000 people, so first-time applicants each year correspond to roughly 2–6 % of a cohort. Figure 2 shows the distribution of applicants' main diagnoses.

In addition to a diagnosis, applicants must submit a rehabilitation plan. This plan is prepared together with medical personnel, social workers, or school-based support staff and specifies the accommodations needed to support school attendance or describes other rehabilitative activities the individual plans to undertake.

Figure A.1 summarizes the application process. Applicants submit their materials to KELA. Each application must document the diagnosis and the required accommodations. Examiners assess the applicant's circumstances, the severity of the disability, and the adequacy of the proposed rehabilitation plan.

Applications are processed within insurance districts, whose boundaries follow regional borders, although their exact configuration changed during the study period. Within each district and language group, applications enter a common queue and are assigned

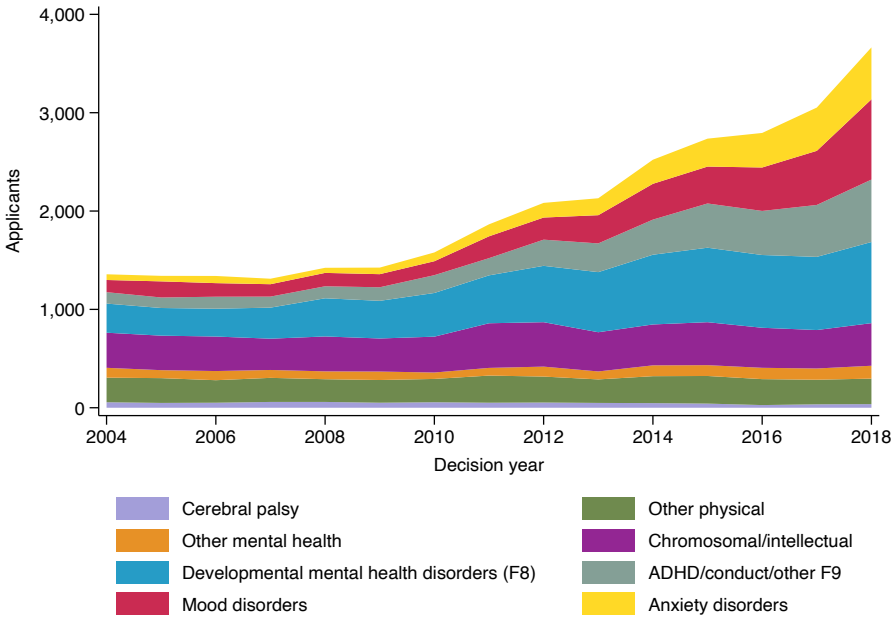


Figure 2: Main diagnosis indicated in the applications across the years 2004–2018.

Notes: The sample includes all applicants, not only approved ones.

to the next available examiner. Applicants cannot choose their examiner. The existence of this system is confirmed by written and verbal communication from KELA, and no written or verbal sources, despite specifically asking about this, have revealed other forms of sorting the applications. As assignment varies by language [6] and district, I account for this variation using language and area-year fixed effects in the empirical analysis (see the methods section).

There are no income or asset tests, either for the individual or the family, and earned income does not reduce the benefit amount.

The benefit is temporary. Once an application is approved, the benefit is paid monthly until one of the following occurs: (i) the recipient graduates from secondary education or completes the approved rehabilitative activity; (ii) the recipient turns 20; or (iii) participation in education or rehabilitation is deemed insufficient. Attendance is monitored, and KELA may discontinue the benefit when participation is insufficient. Grades are not assessed, and graduation is not a condition of receiving the benefit. The monthly benefit was about €700 until 2017 and €900 thereafter [7]. Accepted applicants receive €25,400 on average over the full benefit spell. YRA is taxable. In practice, recipients with no other income pay little tax.

YRA recipients can meet the rehabilitation requirement through several types of activity, although the data do not record which one they undertake. For benefit spells lasting multiple years, the main rehabilitative activity can be assumed to be secondary education. Recipients may also attend other forms of education that do not lead to a secondary education diploma. These studies are not recorded very well in the administrative data, and some individuals attending them have their main activity recorded as *unknown*.

[6] Finland has a large Swedish-speaking minority that uses Swedish in administrative matters.

[7] In 2023 euros.

Among YRA recipients, 59% attend vocational secondary education, 10 % attend general upper-secondary school, 14 % are completing basic education or attending other non-degree programs, and 1 % attend higher education. The remaining 16 % are neither classified primarily as students nor enrolled in secondary education, although some may participate in education not captured by the registers or in other rehabilitation. These shares are close to those reported by Koskenvuo et al. (2014).

2.2 Schooling system in Finland

Finland provides tuition-free comprehensive education from ages 7 to 16. Almost all comprehensive schools are public, and students receive meals and learning materials free of charge. During the study period, compulsory education ended when a student completed comprehensive school or, at the latest, ten years after it began.

Students who enter secondary education choose between an academic and a vocational track. Admission is based primarily on teacher-assigned ninth-grade marks. In the academic track, students follow a standardized curriculum covering mathematics, literature, languages, history, and the natural sciences. Vocational education is organized into occupation-specific tracks that prepare students for specific professions, such as car mechanic, plumber, or practical nurse. While the academic track is primarily intended to prepare students for university studies, both tracks qualify students for higher education. Secondary education typically lasts three years and is tuition-free, as is higher education.

Secondary schools offer a range of support measures for students with learning difficulties and disabilities. Special schools and programs provide more individualized instruction and allow for slower progression, with graduation often occurring after four to five years instead of the standard three. These schools and programs enroll only a small share of students; most special-needs education takes place in regular schools. All special educational support is offered free of charge.

2.3 Healthcare

Finland has a universal public health insurance system that covers the entire population. Healthcare services, particularly for children and adolescents, are predominantly provided by the public sector. The public sector provides almost all treatment for serious conditions and emergencies. Private healthcare is also available, but it is accessed mainly through employer-provided coverage and typically focuses on preventive care or mild conditions. Private healthcare therefore plays a limited role for the population studied in this paper, and the public healthcare registers capture most healthcare my sample receives.

Adults make modest out-of-pocket payments for public healthcare, subject to an annual cap of about €600. A similar annual cap applies to out-of-pocket expenditures on prescription drugs reimbursed by KELA, which include the vast majority of commonly prescribed medications.

Health insurance coverage does not depend on YRA receipt. Copayments in the public system are generally low, although patients often pay a substantial share of psychotherapy costs. Unlike Supplemental Security Income in the United States, YRA does not confer

health insurance eligibility. Any estimated health effects therefore do not operate through a statutory change in insurance coverage.

2.4 Other benefits

Those not receiving YRA can receive multiple alternative benefits. Disability Insurance/Pension (DI) [8] is similar to YRA in that it is based on medical screening, but with a separate application process. Individuals applying need both a diagnosis from a medical doctor and acceptance from an insurance institution. For young individuals with little or no work history, this institution is usually KELA.

For the first few years, DI is awarded as a temporary DI, but if the inability to work persists, it can be made permanent. An individual in DI can have small earnings without the benefit being reduced. Eligibility for DI starts at age 16, and the amount during this period was mostly equal to YRA or somewhat higher.

Applicants may also qualify for unemployment benefits, student aid, housing allowance, or social assistance. Many of these programs are means-tested on personal or household income and provide little support to young people who live with their parents. For instance, student aid for individuals attending secondary education and living with parents is limited only to the families with little or no earnings. Applicants denied YRA therefore tend to receive no other benefits or smaller transfers, usually without a school-attendance requirement.

3 Data

I combine KELA records on YRA applications from 2004–2018 with Finnish administrative registers that follow applicants annually and, where relevant, link them to their parents. Outcome coverage varies across registers, with the longest series running through 2023. Statistics Finland stores the linked data on secure servers, where I conduct all analyses. Information on how to apply for access is available from the author. All euro amounts in the paper, including earnings, benefits, and healthcare costs, are adjusted for inflation to 2023 prices using the consumer price index.

3.1 Youth Rehabilitation Allowance Application data

KELA records all YRA applications and benefit payments.[9]

The application records identify the applicant and examiner and report the decision (approved, rejected, or discontinued), decision date, diagnosis code, and municipality of residence. They cover 2004–2018. Applicant identifiers link to the other registers, but examiner identifiers do not.

Separate payment records describe the payment dates and amounts and whether KELA paid the benefit to the applicant or a parent or guardian. About 98 % of beneficiaries receive payments directly.

[8] The Finnish program is formally a disability pension. I use Disability Insurance (DI) to match the terminology in the international literature.

[9] KELA is the Social Insurance Institution of Finland.

I construct the analysis sample as follows. I use individuals' first application, no matter if they were eventually accepted later. I exclude applicants with missing diagnosis codes—all of whom were rejected—and those older than 20 at their first application. The final sample contains 30,623 applicants. The sample is unbalanced: follow-up differs, individuals leave the panel after death, and some registers do not cover the latest years. Sample sizes therefore vary across outcomes and follow-up periods.

3.2 Labor market and education

Statistics Finland's FOLK modules provide annual labor-market and demographic outcomes. The FOLK PERUS module (Statistics Finland, 2025d) includes yearly information on earnings, main activity, disability insurance receipt, place of residence, and other outcomes for the full Finnish population, including applicants and their parents. These data are mainly available for the years 1987–2023. I classify a person as not in employment, education, or training (NEET) if their main activity during the year is neither employment nor study and they are not completing military service.[10] I use wage income as a measure of earnings.

The FOLK LAPS module includes links between biological parents and children (Statistics Finland, 2025c). Using these links, I am able to match the applicants to their parents. The parents not found in this data are presumably dead, unknown, or have migrated from the country. For these individuals, I use an indicator for missing parents in analyses where this is necessary. Using this linkage, I also identify the siblings of the original applicants for the spillover analysis.

I use two datasets on education. First, I use the TYHR (Statistics Finland, 2025b) dataset to describe the educational history of the individuals. These data are available for the years 1998–2023 and include the GPA of the students graduating from the basic education and their applications to the secondary education. I use whether a student appears in these data as an indicator of graduation and GPA as a measure of academic performance. Since the data do not cover the whole study period, a student not found in the data can also be one who graduated outside the years covered. The EDUC register (Statistics Finland, 2025a) records enrollment in secondary and higher education.

3.3 Healthcare data

I use three Finnish healthcare registers to describe the health and healthcare use of the population. The National Institute for Health and Welfare maintains HILMO, which covers inpatient and specialist outpatient care from 1995 to 2023, and AvoHILMO, which covers public primary care from 2011 to 2023. The registers report visit dates, diagnoses or reasons for care, emergency status, and procedures. Because the public sector provides most serious and emergency care, these registers capture most such events in the sample.

The third register covers reimbursed prescription drugs. KELA reimburses most prescription medicines, usually at the pharmacy, so the register captures most prescribed drug use in Finland.

[10] Finland requires men to complete 6–12 months of military or civilian service, although people with health problems may be exempt.

I use the registers to measure both serious health events and broader healthcare use. Measures of serious health events include hospital days, emergency visits, and inpatient visits with a mental-health diagnosis. Because these events generally require care regardless of cost, they should primarily reflect underlying health. Broader utilization measures include total healthcare spending, prescription purchases, and specialist and primary-care visits. These outcomes may respond to YRA through changes in health or care-seeking. I return to this distinction when interpreting the estimates in the results section.

I define substance-use problems using public healthcare diagnoses for substance use and addiction (ICD-10 codes F10–F19). Where both registers are available, the outcome uses specialist and primary-care records.

HILMO reports provider costs from 2012 onward, although costs remain unavailable for some administrative units. I cap visits at one per person, day, and diagnosis to avoid counting transfers between units as separate visits. Among records with observed costs, I calculate average costs by main ICD-10 diagnosis and length of stay and use these averages to impute costs for all visits. I validate the imputation among visits with observed costs.

The prescription register reports the public reimbursement for each purchase. The primary-care data begin in 2011 but contain no costs, and no official method is available to estimate them from the information recorded. I therefore exclude primary care from the cost calculations.

3.4 Other data

To assess participation in criminal activities, I use both criminal convictions and police reports. The conviction data cover the years 1995–2023 and the police report data the years 2006–2023. Both data sources include dates of the crimes and the type of crime committed. I construct measures of drug-related crime and violent crime from the crime codes.

I use data from the National Enforcement Authority (in Finnish, Ulosottolaitos) to measure financial distress. The data include all enforced debts from the years 2008–2023. A debt or a bill ends up in enforcement after the creditor files for enforcement, usually with a delay of multiple months from the due date. As argued in Maczulskij et al. (2025), ending up in debt enforcement is seldom desirable for the indebted: wages, benefits, and property can be garnished, and individuals are subjected to severe monitoring and other procedures. Unlike, for instance, seeking bankruptcy in the US, it is not a strategic choice, which makes it a good measure of financial distress.

3.5 Descriptive results

Table 1 presents summary statistics for the applicants, separately for those approved and those rejected.

Applicants have poor average labor-market prospects. Figure 3 shows that, at age 25, about 30 % are employed and 15 % are students; the remainder are unemployed or receive DI.

In my main heterogeneity analysis I use an index of labor-market prospects. To estimate the index, I use application-year and diagnosis-group indicators to predict the probability of DI receipt two years after application. While other information could also be used, this

	(1) Gen. pop.	(2) Approved	(3) Rejected
Student	0.976	0.997	0.996
Employed	0.024	0.003	0.004
Distinct admissions	1.02	8.34	6.50
Healthcare costs	559	5,779	4,094
Anxiety diagnosis	0.053	0.091	0.079
Depression diagnosis	0.052	0.098	0.082
Single parent HH	0.211	0.290	0.305
Urban	0.807	0.781	0.804
Mother age	44.04	44.19	44.05
Father age	46.45	47.03	46.69
Mother DI	0.027	0.085	0.075
Father DI	0.040	0.081	0.079
Mother married	0.697	0.558	0.562
Father married	0.691	0.550	0.553
Mother employed	0.814	0.699	0.721
Father employed	0.831	0.727	0.740
Mother wage income	24,008	18,994	19,697
Father wage income	33,096	26,360	26,742
Accepted (1st appl.)	0	1	0
Ever in YRA	0	0.998	0.266
Total YRA benefits	0	22,967	5,397
Observations	1,476,350	20,832	9,735

Table 1: Summary statistics for the applicants, by approval status.

Notes: All variables are measured in the year before the first application. Euro amounts are in 2023 prices. Gen. pop. is the sample of all other individuals aged 15.

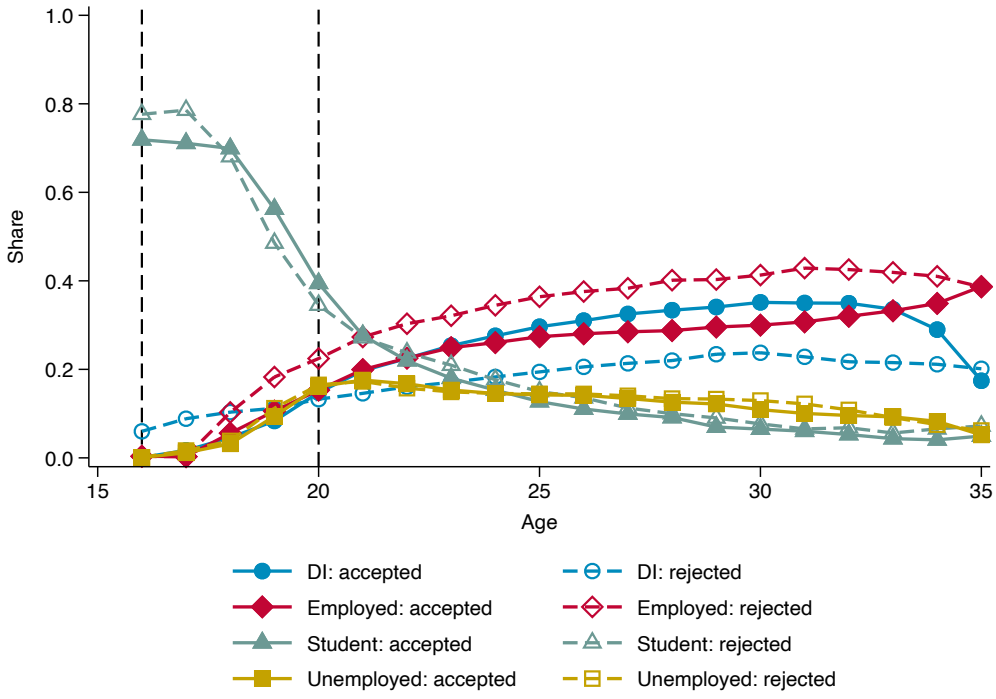


Figure 3: Shares in each main activity by age and YRA approval status.

Notes: The figure excludes military service and uses an unbalanced panel. Vertical lines mark the main YRA eligibility period.

is transparent since the prediction is only based on "severity of the disability". To limit overfitting, I construct out-of-fold predictions using two-fold cross-fitting and train the prediction model only on rejected applicants. I randomly split rejected applicants into two folds. I estimate a regression forest on one fold and predict two-year DI receipt in the held-out fold, reverse the training and prediction samples, and combine the resulting out-of-fold predictions. I then use the fitted models to construct predictions for approved applicants. This avoids incorporating the effects of treatment into the index.

The results that use this index are similar if a different time horizon or different outcome for the prediction is used. As can be seen in Fig. 4, the labor-market-prospects index is strongly correlated with the share that is outside of the labor force later in adulthood, showing that it captures labor market prospects.

I classify applicants in the top quintile of the index as having low labor-market prospects and the remaining applicants as having higher prospects within the applicant sample. A descriptive table for this group is shown in Table A.7. Both groups nevertheless have worse health and socioeconomic backgrounds than the general population.

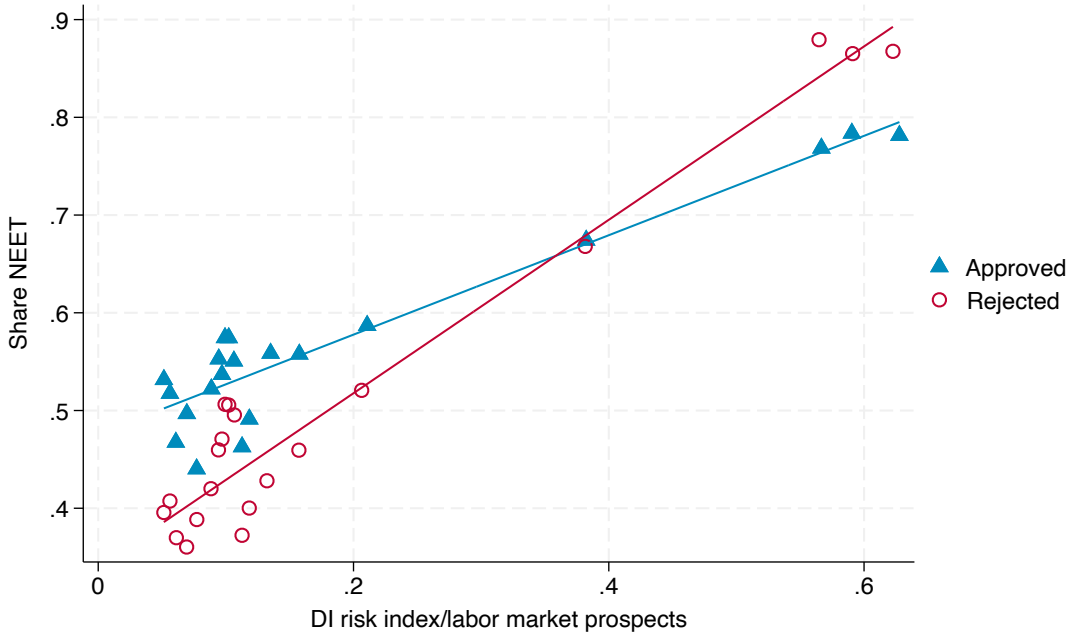


Figure 4: Adult NEET (not in education, employment or training) rates by predicted labor-market prospects and YRA approval status.

Notes: The figure plots mean NEET rates in 20 equal-sized bins separately for approved and rejected applicants. Higher values of the index indicate worse labor-market prospects.

4 Methods

4.1 Target populations and parameters

To estimate the effects of YRA, we can start by considering the simple comparison of approved and rejected applicants

$$Y_{it} = \beta_0 + \beta_1 Approved_i + \beta_2 X_i + \epsilon_{it} \tag{1}$$

where i indexes applicants, t indexes time, $Approved_i$ indicates first-application approval, and X_i contains covariates observed before the decision. This strategy relies on the assumption of conditional ignorability: conditional on X_i, approval is independent of the potential outcomes,

$$\big(Y_{it}(1), Y_{it}(0)\big) \perp Approved_i \mid X_i.$$

Approved and rejected applicants differ on many observed background variables, even within the applicant sample (Table 4). This suggests the presence of unobserved factors correlated with approval. My main design exploits the quasi-random assignment of applications to examiners who differ in their approval rates (Section 4.2). I also report estimates that adjust for selection on observed covariates.

Before describing the estimators, I define the parameters they estimate. Treatment is approval of the first application.

Treatment effects likely differ across applicants. Applicants can vary substantially in their labor market prospects. For those with relatively strong prospects, approval may mainly reduce work while they receive YRA. For those with poor prospects, the alternative to YRA may be Disability Insurance rather than employment. Different averages therefore answer different questions.

The average treatment effect (ATE) over all applicants, $ATE = E[Y_{it}(1) - Y_{it}(0)]$, is the relevant parameter for large changes to the program, such as approving nearly all or hardly any applications. Examiner IV instead estimates a weighted average of the effects on compliers: applicants whose approval depends on their assigned examiner. This local average treatment effect (LATE) is directly relevant for changes to the approval threshold.

The LATE does not describe all applicants. The compliers for the most part have average or better than average labor market prospects (Section 4.2.3). The effects on applicants with the poorest prospects are also of interest. These effects may differ from those on marginal applicants and therefore require going beyond the IV.

Examiner IV is the primary design. I use selection on observables for populations and outcomes the IV does not cover well: the full applicant population, applicants with the poorest labor market prospects, and outcomes for which the IV estimates are imprecise. Where the strategies estimate the same parameter for the same population, I compare their estimates. Similar estimates support the assumptions of the selection-on-observables design. Table 2 summarizes the designs.

4.2 Examiner IV

An application arrives at the insurance district covering the applicant's municipality and is handed to the next examiner in the queue. Within region, year, and language, applicants receive different examiners because of their position in the queue. Examiners differ in their approval rates, and this variation provides the instrument for first-application approval.[11]

[11]The design has been used in many papers on the economics of crime (Kling, 2006), foster care (Doyle Jr, 2007; Bald et al., 2022; Helénsdotter, 2026), and disability insurance (Maestas et al., 2013; French and Song, 2014; Kostøl and Mogstad, 2015; Autor et al., 2019), among others.

Design	Identifying assumption	Estimand	Population	Used for
Examiner IV	Relevance, conditional random assignment, exclusion, average monotonicity, and extensive-margin compliers only (A1–A4)	LATE	Compliers (about 23% of applicants), concentrated in the middle of the distribution of labor market prospects	The main short-run and long-run estimates
Selection on observables (DDML)	Conditional ignorability given the covariates in Section 4.3, and overlap	ATE, also by subgroup	All applicants, and subgroups defined by predicted DI risk; reweighted to the compliers for comparison with the IV	Effects for the full applicant population, for the applicants with the lowest labor market prospects, and for outcomes where the IV is imprecise

Table 2: Empirical designs.

The first stage relates approval to the leniency of the examiner (j):

$$Approved_i = \alpha z_{ij} + \mu_c + \eta_i. \tag{2}$$

Here, z_{ij} measures the examiner's tendency to approve applications, and μ_c denotes the time and place fixed effects needed to make examiner assignment as good as random.

The second-stage equation is

$$Y_{it} = \beta_0 + \beta_1 \widehat{Approved}_i + \mu_c + \epsilon_{it}. \tag{3}$$

In most specifications, I include only the fully non-parametric fixed effects. Because I include the necessary controls in fully saturated form, the specification satisfies the *rich covariates* condition of Blandhol et al. (2026). The 2SLS estimand is therefore weakly causal rather than a combination of treatment effects that can carry negative weights.

I cluster standard errors at the applicant level in pooled regressions to account for repeated observations. I do not cluster at the examiner level because assignment occurs at the applicant level, in line with the previous literature (Chyn et al., 2024). The endogenous variable is approval of the first application.

I use first-application approval as the main treatment because it is most likely to satisfy the exclusion restriction. Around 20% of rejected applicants are approved on a later application, and the instrument also raises the probability of ever being approved (Table 3). Using *ever approved* as the endogenous variable changes the scale of the estimates but otherwise yields similar results (Fig. A.10).

I construct residualized leave-one-out examiner leniency, following Agan et al. (2023). Leaving out an applicant's own case breaks the finite-sample correlation between the instrument and that applicant's decision. Assignment takes place among examiners handling applications in the same region, year, and language group. I therefore residualize approval decisions using region × year × language fixed effects. I call these the *necessary controls* because random assignment holds only conditional on them.

In the Appendix, I check alternative controls and residualization choices. The estimates are similar when, for example, I add the month of decision to the equations (Fig. A.16).

The instrument is

$$z_{ij} = \frac{1}{N_j - 1} \times \left(\sum_{k=1}^{N_j} (Approved^*_{jk}) - Approved^*_i \right). \tag{4}$$

Here, N_j is the number of cases handled by examiner j, and $Approved^*$ is the residual from regressing approval decisions on region × year × language indicators. Subtracting $Approved^*_i$ removes the application's own contribution and gives the leave-one-out mean. Each application therefore receives a different value of the instrument. I calculate the main instruments once with the full application sample and use these throughout.

The sample contains 379 examiners, who handle 79 cases on average. Appendix Fig. A.17 shows that the estimates are similar under alternative minimum-caseload restrictions.

4.2.1 Empirical Bayes shrinkage

Examiner IV studies usually use either residualized leave-one-out examiner leniency or jackknife IV. Both estimate each examiner's leniency from that examiner's own caseload. These caseloads are small: the average examiner in my data handles 79 applications, and many handle far fewer. The resulting leave-one-out mean is noisy. This noise does not threaten validity, but it weakens the first stage and widens the second-stage confidence intervals.

To reduce this noise, I use an empirical Bayes approach. The observed leave-one-out mean z_{ij} is the sum of the examiner's true leniency and estimation error. The procedure shrinks noisier estimates further toward the average:

$$z^{EB}_{ij} = \bar{z} + \lambda_j (z_{ij} - \bar{z}), \qquad \lambda_j = \frac{\sigma^2_z}{\sigma^2_z + s^2_j}, \qquad s^2_j = \frac{\hat{\sigma}^2_j}{N_j} \tag{5}$$

where $\hat{\sigma}^2_j$ is the within-examiner sample variance of approval decisions residualized on region × year × language fixed effects. The shrinkage weights use $s^2_j = \hat{\sigma}^2_j / N_j$, the estimated sampling variance of the full-caseload mean, and are applied to the leave-one-out instrument z_{ij}. I estimate σ^2_z as the unweighted sample variance of examiners' full-caseload residual means minus the unweighted average of s^2_j, setting negative estimates to zero. The shrinkage target $\bar{z}$ is the unweighted mean of these examiner means. Appendix Section A.6.6 gives the formulas.

For a given within-examiner variance, smaller caseloads receive more shrinkage toward the average. Estimates for examiners with large caseloads remain close to their leave-one-out means. The procedure does not add identifying assumptions beyond A1–A4.

Empirical Bayes shrinkage is common in settings such as teacher value added (Chetty et al., 2014), but less common in examiner IV designs. Humphries et al. (2025) use it as a check against measurement error in the instrument. Frandsen et al. (2026) develop the approach concurrently with my work. They establish its theoretical properties, derive optimal shrinkage formulas, and provide simulation and empirical evidence. My choices are close to theirs, and their paper contains the proofs.

In my application, shrinkage leaves the point estimates similar but reduces standard

errors for the main outcomes by around 25% relative to the standard leave-one-out mean instruments (Fig. A.16). This is a combination of being able to include the examiners with small amounts of cases and improving the first stage in general.

Appendix Section A.6.6 provides additional details and simulation evidence on empirical Bayes IV. Unless stated otherwise, my IV results use shrunk instruments because they provide more precise estimates. Section 5.1.4 shows that the point estimates are similar under empirical Bayes and the standard leave-one-out procedure.

4.2.2 Identifying assumptions

I use four assumptions to interpret the IV estimate as a LATE for applicants whose approval responds to examiner leniency:

A1. Relevance. The instrument must have a strong effect on the endogenous variable, that is, approval of the application.

A2. Independence. The instrument is as good as randomly assigned, conditional on the necessary controls.

A3. Exclusion. The instrument only affects outcomes through treatment.

A4. Average monotonicity. More lenient examiners are more likely to approve an application on average, though this need not hold for every applicant.

Under these assumptions, the IV estimator identifies the average effect of having the first application approved for the compliers (LATE).

A1 Instrument Relevance

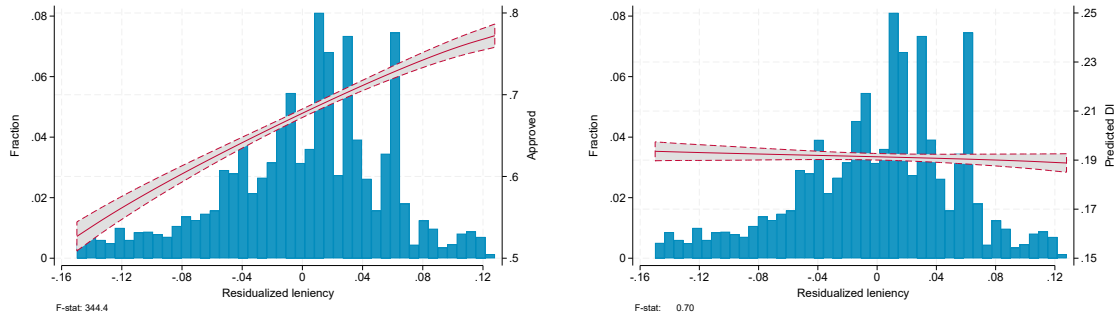
Examiner leniency strongly predicts approval. The histogram in Figure 5a displays the distribution of the instrument, with fractions in each bin on the left y-axis. The line displays a linear local polynomial (with bandwidth 0.05) of the instrument z_{ij} on first-application approval. A one SD increase in the instrument raises the approval probability by 4.9 pp, or 7.2% of the 68% mean. This also translates to increases in amount received and probability of receiving any YRA (Fig. A.6).

Table 3 shows that the first stage is strong: the coefficient on examiner leniency is 0.803, and the (Kleibergen-Paap) first-stage F-statistic in the main specification is 344. F-statistics in leniency designs are not comparable to those for a single, one-dimensional instrument, so the $F = 10$ rule of thumb is not the right benchmark (Hull, 2017; Chyn et al., 2024). Following Andrews et al. (2019), I also report Anderson–Rubin p-values for the main results (Table 6), which allow inference robust to weak instruments.

A2 Instrument exogeneity

Within region, year, and language, examiner assignment must be independent of applicants' potential outcomes. Applications are assigned from a running queue. I residualize approval decisions using region × year × language fixed effects and include the same fixed effects in the regressions. Written and oral communication from KELA confirms the queue system, and I found no evidence of additional sorting.

To my knowledge, Reiterä (2021) first uses this design in the YRA setting to study active rehabilitation policies. They also argue that exogeneity is plausible after including



(a) Application approval

(b) Predicted DI risk

Figure 5: Application approval and predicted DI risk by examiner leniency.

Notes: The histograms show the distribution of the instrument z_{ij} in the estimation sample (left y-axis). The lines show linear local polynomials of first-application approval (panel a) and predicted DI risk (panel b) on residualized examiner leniency (right y-axis). Panel b includes 95% confidence intervals. Constructed using pretreatment variables, mainly health and family background, to predict DI receipt 5 years after application.

	(1) First application approv.	(2) Ever in YRA	(3) Total YRA Benefits
EB IV	0.803***	0.547***	13792.6***
	(0.0433)	(0.0395)	(1330.2)
N	30623	30623	30623
F-Statistic	344.4	191.7	107.5

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Table 3: First-stage estimates for alternative treatment measures.

Notes: "Ever in YRA" is the probability of receiving YRA at any point in the observation period. "Total YRA benefits" is total YRA receipt over the observation period. All regressions use the same instrument; it is not estimated again for each outcome. All specifications include the necessary region × year × language fixed effects used to construct the instrument.

similar controls, but provide little institutional or empirical evidence on possible threats to random assignment.

Random assignment implies that examiner leniency should be unrelated to applicants' pretreatment characteristics. Figure 5b checks this implication using predicted DI risk. The index combines pretreatment measures of socioeconomic status and healthcare use that predict later outcomes.[12]

The curve is flat: examiner leniency is not meaningfully correlated with predicted DI risk. Scaling this association by the correlation between the risk index and the key outcomes implies a bias of around 0.5 pp in post-benefit employment. This is less than 5% of the IV estimate.[13] The small association supports the assumption that examiners are randomly assigned.

Table 4 tests joint balance across observable characteristics. Approval is strongly related to applicants' backgrounds, especially their health measures. In contrast, the coefficients relating these characteristics to examiner leniency are small: two of 30 are significant at the 10% level, and I cannot reject that they are jointly zero ($p = 0.62$).

[12] Any imbalance matters for bias only if the unbalanced characteristic also predicts the outcome.

[13] The implied upper confidence limit for the bias is around 1.1 pp.

Two further checks support the assignment mechanism. For health outcomes observed from birth onward, I estimate placebo effects before applicants are eligible for YRA and present these together with the estimates for the periods after application. Random assignment implies no relationship in either levels or trends before application, and the estimates are consistent with zero. Table A.11 also shows that the first-stage coefficient changes little when I add controls beyond the necessary fixed effects.

These checks cannot rule out unobserved sorting. If sorting occurs, it would most likely operate through information available at the application stage. In robustness checks, I add decision month, diagnosis category, parental socioeconomic background, and ninth-grade GPA as control variables. The estimates change very little. For unobserved sorting to explain the results, it would have to be large relative to sorting on these observed characteristics. To account for the possibility that some areas would systematically break the random assignment, I also leave out areas and estimate the main effects again. The results remain similar, which reduces these concerns.

A3 Exclusion restriction

Random assignment identifies the reduced-form effect of being assigned a stricter examiner. Identifying the effect of first-application approval also requires examiners to affect outcomes only through that decision. This exclusion restriction can fail in settings such as criminal justice, where judges often make decisions about several treatments (Humphries et al., 2025).

The exclusion restriction is plausible here because examiners only approve or reject applications based on the submitted information. The decision is communicated through KELA. Examiners should usually neither advise applicants about other benefits nor meet them in person or have extensive communication through other means. They do not make decisions on other benefits or policies. The institutional details make the case for exclusion restriction a strong one in my setting.

A4 Monotonicity

Strict monotonicity requires a more lenient examiner to approve every application that a stricter examiner would approve (Imbens and Angrist, 1994). This is demanding in examiner IV designs. It fails, for example, if some examiners are stricter toward mental health conditions but more lenient toward other conditions. Frandsen et al. (2023) and Chyn et al. (2024) call this *strict monotonicity*. Under strict monotonicity, researchers can also estimate marginal treatment effects (Mogstad et al., 2018).

My main results rely on the weaker *average monotonicity* assumption. Applicants may respond differently to examiners, but more lenient examiners must increase approval on average. The IV estimates can then be interpreted as "convex averages of the treatment effects among compliers" (Frandsen et al., 2023). Strict monotonicity would also be needed to estimate marginal treatment effects.

I assess average monotonicity in three ways, described in Appendix Section A.3.1. The first stage is positive and significant in every subsample (Maestas et al., 2013) and in each split by illness category and gender (Bhuller et al., 2020). The binary lenient-against-strict

	(1) Approved	(2) EB IV
Applicant		
Month of dec.	-0.00433***	-0.0000106
	(0.000766)	(0.000104)
Urban	-0.0114*	-0.00130
	(0.00657)	(0.000888)
Male	0.0240***	0.000786
	(0.00551)	(0.000745)
Student	-0.0205**	-0.00133
	(0.00840)	(0.00114)
GPA (9th grade)	-0.0219***	0.000104
	(0.00228)	(0.000308)
Age	-0.0470***	-0.000636
	(0.00337)	(0.000455)
Parents		
Father age	0.000884	0.0000557
	(0.000554)	(0.0000750)
Mother age	-0.0000544	-0.00000947
	(0.000620)	(0.0000838)
Father wage inc.	0.000117	0.0000113
	(0.000138)	(0.0000187)
Mother wage inc.	-0.000235	-0.0000463*
	(0.000203)	(0.0000274)
Father employed	-0.00189	-0.00154
	(0.00716)	(0.000968)
Mother employed	-0.0104	0.000648
	(0.00739)	(0.001000)
Mother gen. HS	-0.00281	-0.000608
	(0.00629)	(0.000850)
Father gen. HS	0.000239	0.000878
	(0.00769)	(0.00104)
Mother DI	0.000363	0.00147
	(0.00984)	(0.00133)
Father DI	-0.00886	-0.00240*
	(0.0102)	(0.00138)
Mother married	0.00461	-0.000312
	(0.00540)	(0.000730)
Diagnosis in application		
Mood/Anx/Dev. dis.	-0.0841***	0.0000300
	(0.00774)	(0.00105)
ADHD/conduct	-0.117***	-0.00000438
	(0.0101)	(0.00137)
Other MH	-0.0368***	0.00119
	(0.0141)	(0.00190)
Other Physical	-0.146***	-0.0000604
	(0.00948)	(0.00128)
Healthcare use		
Hosp. visits	0.00129***	0.0000173
	(0.000199)	(0.0000268)
Any Depression vis.	0.0104	-0.000875
	(0.00853)	(0.00115)
Any anxiety vis.	0.0138	-0.00108
	(0.00843)	(0.00114)
Any intell. dis. visit	0.0722***	0.000783
	(0.00777)	(0.00105)
Any ADHD/conduct visit	0.0455***	-0.000937
	(0.00791)	(0.00107)
Any nervous system vis.	0.0293***	-0.000322
	(0.00956)	(0.00129)
N of diff. prescr.	-0.00114**	-0.00000734
	(0.000487)	(0.0000658)
N of antidepr. prescr.	0.00348	-0.000327
	(0.00798)	(0.00108)
N of Diabetes prescr.	-0.0134	0.000166
	(0.0124)	(0.00168)
Observations	30623	30623
F	52.56	0.899
p	3.24e-305	0.624

Standard errors in parentheses
* p<0.1, ** p<0.05, *** p<0.01

Table 4: Balance tests.

Notes: Column 1 regresses first-application approval on applicant characteristics; column 2 regresses examiner leniency on the same characteristics. The bottom rows report joint F-tests and their p-values. Standard errors are in parentheses.

instrument proposed by Sigstad (2024) also gives estimates similar to the main results (Fig. A.16).

4.2.3 Compliers

If assumptions A1–A4 hold, the IV estimates the average treatment effect for compliers. These applicants are approved by some examiners but rejected by others. Compliers likely fall in the middle of the distribution of labor market prospects: applicants with the poorest prospects may be approved by every examiner, while those with the strongest prospects may be rejected by every examiner. Although individual compliers cannot be identified, I can describe the group in several ways.

Following Maestas et al. (2013), under strict monotonicity the share of compliers can be estimated from Fig. 5a or as $\alpha \cdot (\bar{z_j} - \underline{z_j})$, where $\bar{z_j}$ and $\underline{z_j}$ are the highest and lowest values of the instrument. Under average monotonicity, this expression instead gives a lower bound on the complier share (Norris et al., 2021). Compliers account for roughly 23% of applicants, so the LATE is relevant for a substantial share of the applicant population.[14]

The estimates nevertheless remain local to the complier population. Without additional assumptions, they are not informative about treatment effects for applicants whom every examiner would approve or every examiner would reject.

Following Bhuller et al. (2020), I use complier weights to compare DDML and IV estimates for comparable populations (Table 11). These weights are constructed by a comparison of first-stage coefficients in different subsamples.

Although individual compliers are unobserved, Frandsen et al. (2023) show how to estimate their mean characteristics by simple regression. I report these estimates in Table 8. Compliers are less likely than the average applicant to have disabilities severe enough to qualify for adult Disability Insurance.

4.2.4 Marginal Treatment Effects

A way of inspecting heterogeneous treatment effects and understanding the external validity of the IV estimates is to estimate Marginal Treatment Effects (MTEs) (Heckman and Vytlacil, 2007; Carneiro et al., 2010). The approach has attractive features: it takes into account the information available in the instrument in more detail and this can be used to estimate different parameters, even the Average Treatment Effect under strict assumptions.

The MTE approach has some drawbacks in my setting. First, the estimation and interpretation as usual requires *strict monotonicity* (Frandsen et al., 2023), instead of the average monotonicity I invoke in my setting. This assumption imposes rather implausible assumptions on the examiner behaviour (Mogstad and Torgovitsky, 2024). While it is possible that the MTEs can be regardless estimated and interpreted in a different way under violations of strict monotonicity (Sigstad, 2024), the literature on this is currently not settled.

Secondly, the MTE does not provide a clear way to estimate the effects on the large group of always-takers in my setting, especially the interesting group of individuals with low

[14] The formula is $\hat{\alpha} \cdot (\bar{z_j} - \underline{z_j})$, evaluated at the first-stage coefficient $\hat{\alpha} = 0.803$ and at the highest and lowest values the instrument takes in the estimation sample.

labor market prospects. The unobserved resistance does not directly map on any observable feature that could be used in understanding the mechanisms or targeting, and the results on the group that are most likely to be approved to the benefit are noisy and sensitive to the parametric assumptions made, as the instrument does not span the entire interval.

With these caveats in mind, I report MTE estimates in the Appendix as suggestive evidence. The drawbacks of the approach in my setting also motivate another way of understanding the effect of the Youth Rehabilitation Allowance on groups beyond the marginal applicants: comparison of observably similar rejected and approved applicants.

4.3 Selection-on-observables approach

As mentioned in Section 4.1, a possible assumption is ***the conditional ignorability assumption***:

$$\big(Y_{it}(1),\, Y_{it}(0)\big) \perp Approved_i \mid X_i, \qquad 0 < P(Approved_i = 1 \mid X_i) < 1. \tag{6}$$

The first part says that among applicants with the same observed characteristics, approval is unrelated to what would have happened to them under either decision. The second part, overlap, says that no combination of characteristics is either always approved or always rejected. This assumption would mean that an adequate selection-on-observables strategy would yield unbiased estimates of treatment parameters.

A common perception since the seminal paper of LaLonde (1986) has been that when evaluating public policy programs, covariate-adjustment methods are hence unreliable, as individuals select into programs based on unobservable variables. Literature has pointed out that selection-on-observables methods have a hard time replicating the experimental benchmarks (LaLonde, 1986; Smith and Todd, 2005). However, with the increased use of rich administrative data and modern estimators, the case for selection-on-observables approaches is not as hopeless as once thought (Dehejia and Wahba, 2002; Imbens and Xu, 2025). By ensuring overlap, meaning that individuals with similar probabilities of treatment are compared, and using more detailed background data, experimental benchmarks can often be recovered. This is not a guarantee that such a strategy could work in my setting, but I argue that multiple details make this approach attractive.

My data allow using large amounts of background information to compare plausibly similar individuals and then verifying these estimates using placebo outcomes and balance tests measured prior to the treatment. My setting is an example of this: I have a natural control group of rejected applicants, plausibly more similar to the approved applicants than to the general population. I observe detailed data on the individuals, including much of the data available on the applications and a long history of healthcare use, allowing me to approximate the information observed by the examiners. As an additional benefit, I can compare my estimates (weighted to the complier population) to the IV estimates, providing an additional layer of validation.

The stability of the coefficients to additional covariates can also be assessed in the spirit of Altonji et al. (2005) and Oster (2019). I also observe many outcomes, like the healthcare use and parental labor market ones, prior to treatment. The treatment cannot have an effect on these outcomes, and any viable method should find a null effect here (Imbens

and Rubin, 2015). These checks provide a solid basis for use of selection-on-observables methods. Still, the results rely on strong assumptions, which should be acknowledged.

Against the weight of the conditional ignorability assumption, the benefits of assuming conditional ignorability should be considered. There are two main benefits. While the IV estimates are sufficiently precise to establish economically important effects on adult employment and NEET status, they are less informative for some other outcomes like health and crime and cannot distinguish plausible effects among the highest-risk applicants, for whom examiner leniency generates much less approval variation.

This problem is not unique to my paper but shared with many papers using an (examiner) IV design. Secondly, the IV compliers are a selected group of the applicants, and extrapolating the effects outside of these compliers requires assumptions that are also not innocent. The problems are related in that the IV design has especially low power on some interesting subgroups that are under-represented in the complier population, for instance, the individuals with the most severe disabilities. For this group, the IV estimate on employment is -0.07 (SE $= 0.17$, F-statistic $= 24$, using similar choices as in Table 6). At 80% power and a 5% two-sided test, the minimum detectable effect is $\approx 2.8 \times 0.17 = 0.48$, more than three times the untreated employment rate of 15% in this group. Even an effect that eliminated employment entirely would be too small for the IV to detect. Usual approaches to extrapolating the LATE from IV designs without full support to the whole sample, like parametric MTE, rely on strong assumptions about the functional form as well as strong monotonicity. In addition, these approaches will also suffer from the power issues inherent with the IV.

More precisely, as the main selection-on-observables approach, I use Double/Debiased Machine Learning (DDML) (Chernozhukov et al., 2018; Ahrens et al., 2024). Relative to conventional selection-on-observables estimators, double machine learning is particularly well suited to settings like mine with a large number of pre-treatment covariates and used in recent empirical work (Chetty et al., 2026). It permits flexible, regularized estimation of the treatment and outcome models using different machine learning methods while using orthogonalization and cross-fitting to reduce overfitting and regularization bias in the treatment-effect estimate. I use an interacted model with logistic regression and gradient-boosted trees in my main estimation. In practice, different estimators give similar estimates when the most important predictors of treatment are included. Standard errors come from the influence function of the DDML estimator and are valid under the conditions in Chernozhukov et al. (2018).

Covariates used in the main estimation are (in total 356, and their interactions):

- **Applicant information:** Gender, Age, indicator for having 9th grade GPA, 9th grade GPA, applicant region, applicant lives in an urban area
- **Application information:** Year and month of application, diagnosis code group $\times$ year of application
- **Parental/family characteristics:** Parent ages, Parent education (3 categories), Parents employed, Parent's wage earnings, Parents in Disability Insurance
- **Healthcare use:** Indicators for any visit and number of visits for the ICD10 main

categories + a more precise categorization for mental health disorders, antidepressant prescriptions filled, stimulant prescriptions filled, total drug expenses, and total hospital days

As the covariates represent much of the information plausibly available on the applications and used by the examiners, they plausibly capture meaningful parts of the selection process. As argued by Angrist (1998) and Lindquist and Santavirta (2014), if we understand the selection criteria and have data on the variables that determine the selection, we can get a good handle on the selection bias in this setting. The most important correlates of approval are the disability categories and previous educational performance. Within disability categories, individuals can yet have differently severe health problems. The measures of healthcare use plausibly capture much of this severity. Finally, parental background can proxy for the severity of the difficulties that applicants have or their skill in drafting the application. In practice, most parental characteristics have little correlation with approval to the program.

It is possible that some combination of, e.g., SES and disability category would affect approval in a particular way. The Double Machine Learning approach is a way to take this into account flexibly without overfitting: both the propensity score and the treatment effect are estimated using cross-folds. When necessary, I cluster the estimation of these at the individual level, so e.g. an individual's earlier observation is not used in the estimation of their propensity score.

4.3.1 Validity checks

Overlap means that the control and treatment groups have similar support in the covariate distributions. Ensuring overlap is important for any identification strategy relying on controlling observables. The most informative covariate is the propensity score, which is a combination of the covariates. Fig. 6 displays the propensity score distributions for the approved and rejected applicants. We can see that without any trimming, the propensity score distributions overlap pretty well. I drop the top and bottom 1% of the propensity scores. A similar graph, but for the high DI risk subgroup can be found in the Appendix Fig. A.20.

Second, we can check whether, when using the chosen strategy, the treatment has an effect on placebo outcomes (measured before the treatment) or immutable characteristics. For this exercise, I remove the chosen covariates (and any derived from them or very correlated with them, like employment and earnings) from the covariate set. I then regress the variables individually on the treatment. Fig. 7 portrays the results from this exercise. While the OLS estimates without controls show that the approved applicants fare worse than the rejected in most characteristics, most of the coefficients from the DDML procedure are not statistically significant and are small also in magnitude. All of the coefficients displayed here and more are included in the estimation of the effects, so even if there were imbalances here it would not necessarily be damning for the design, as they are still controlled for in the results.

I perform a similar exercise separately for the high DI risk subgroup and find similarly mostly small imbalances (Fig. A.21). While some coefficients are significantly different

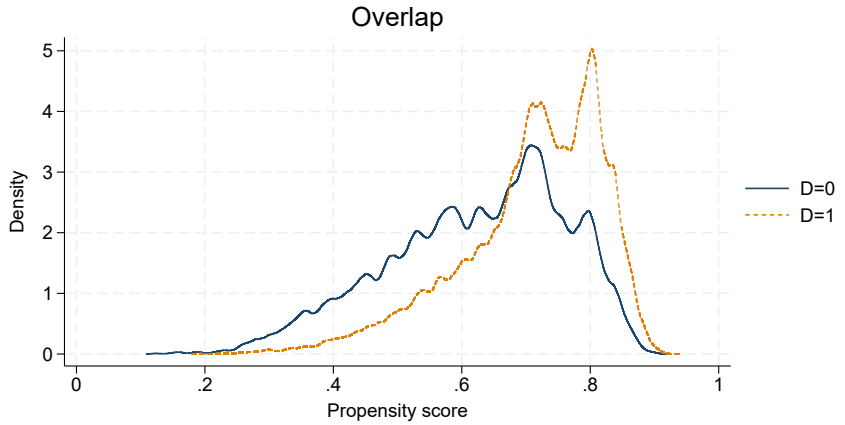


Figure 6: Overlap figure for the covariate adjustment method.

Notes: The figure displays the distribution of the propensity scores estimated by the DDML procedure, using all covariates. 'D=1' refers to the distribution of the treated (approved) individuals, while 'D=0' refers to the distribution of the control (not approved) observations. Note that this propensity score is estimated for the main sample of individuals observed during years 6-11 from first application, for other samples, the figure would look slightly different. A similar figure for the low labor market prospects/high DI risk subsample is available in Fig. A.20.

from zero, there is no particular direction for these biases and even the covariates with any imbalance are included in the covariate set used for estimation.

The validity checks presented here support the conditional ignorability assumption. In the results section, I present additional validity assessments. Stability of the results when including different sets of covariates is intuitively an indication of the extent of selection bias in a setting (Altonji et al., 2005). More formally, I also calculate the δ coefficients proposed by Oster (2019), setting the maximum R^2 to 1.3 times the R^2 of the controlled regression [15]. The Oster δ shows how much more important selection on unobservables would have to be compared to the selection on observables to make the estimates equal to 0. A result with $\delta > 1$ is often considered a robust one: the selection on unobservables would have to be stronger than the selection on observables. I recover values between 2.6 and 6 for my key outcomes (Employment and NEET status). For the especially interesting subgroup of low labor market prospect individuals, adding observed controls moves the estimates away from zero, consistent with selection on observables attenuating the effects. If selection on unobservables operates in the same direction, the employment effect would be more positive than the reported estimate.

A further advantage in my setting is that under the assumptions A1-A4 I have a benchmark I can use to assess the performance of the covariate adjustment in removing selection bias: the IV estimates. A complication is that the IV estimates represent a treatment effect on the instrument compliers only. A simple comparison of the coefficients is therefore not informative of selection bias in the presence of heterogeneous treatment effects. I solve this issue by calculating complier weights using the approach used by Bhuller

[15] Examples of the use of this logic in published papers include Nunn and Wantchekon (2011); Lindquist and Santavirta (2014); Mian et al. (2023); Dustmann et al. (2023) and many more.

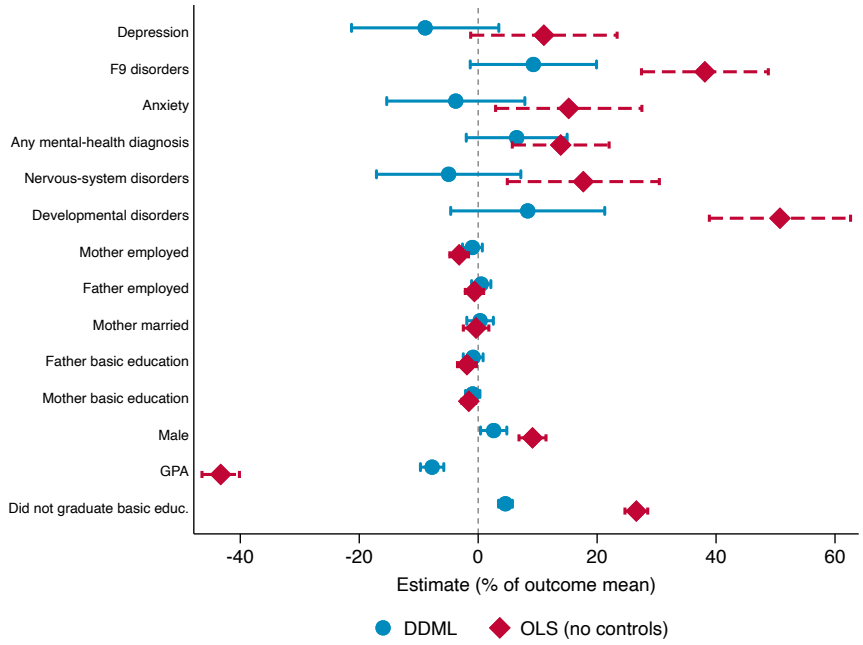


Figure 7: Balance test for the covariate adjustment design.

Notes: For each coefficient, I run the DDML procedure with the listed pretreatment or time-invariant characteristic as the outcome. I leave out from each regression the closely correlated variables and the ones directly derived from this variable. The red markers display the raw associations between approval and a characteristic.

et al. (2020) and weighting my DDML sample with these weights [16]. The results obtained this way are very close to the IV results for the main outcomes.

The fact that adjusting for covariates can remove selection among the compliers is not a guarantee that the same covariates remove selection in other parts of the sample. For this validity check to be informative about the selection in other subgroups, the covariates that close the gap between approved and rejected applicants in the middle of the labor market prospect distribution should also close it at the bottom of it: that selection runs through the same observed variables for the severely disabled as for the marginal applicant, in a similar manner to Aizer et al. (2024). Together with the other validity checks, though, it provides good reasons to find the conditional ignorability assumption a plausible one.

5 Results

YRA approval reduces labor-market participation after benefit eligibility ends, despite having little effect on educational attainment. I first report short-run effects on education, work, and benefit receipt, which help interpret the longer-run results. I then examine health and healthcare use, test the sensitivity of the IV estimates, describe the compliers, and

[16] In practice, this works by splitting the sample and comparing the first stage in these samples. The larger the first stage, the larger the share of compliers in this split of the sample. This restriction does not identify individual compliers; it selects the observable groups in which the first stage indicates the highest concentration of compliers. In the DDML procedure I restrict to the top 20 % of the sample with the highest complier likelihood. The methodology does allow for continuous weights, I instead approximate this for the purposes of my validation exercise. The results are not sensitive to the exact definition.

present results based on covariate adjustment.

5.1 IV results

5.1.1 Short-run impacts

Among marginal applicants, YRA approval generates little additional schooling but appears to displace earnings during benefit eligibility. Table 5 presents cumulative outcomes through age 21, when the education or rehabilitation condition ends. These measures summarize outcomes over the benefit period. I show the main results in panel A, results with additional controls are shown as a robustness exercise.

YRA receipt is conditional on attendance in secondary education or another rehabilitative activity. However, it is still possible that most marginal applicants would attend secondary education even without YRA, leaving little scope for the benefit to raise enrollment. The IV estimates confirm this: years spent as a student barely increase (Table 5). The confidence intervals rule out gains as large as those found for conditional cash transfers in developing countries (Schultz, 2004; Glewwe and Kassouf, 2012); the estimates are closer to those for the UK Education Maintenance Allowance (Britton et al., 2025).

YRA approval raises secondary education completion by around 2 percentage points by age 21, although the estimate is imprecise. As rejected applicants catch up, the completion gap narrows by age 25. Recipients are no more likely to enter higher education afterwards (Fig. A.2), from also a low base. This is in line with the idea that the definition of success in the labor market for these individuals is different from that of the general population.

Somewhat in line with the intentions of the policymakers, YRA approval also reduces Disability Insurance (DI) receipt. This effect is mechanical because YRA replaces DI. Its magnitude indicates how often approved applicants would otherwise apply for and be awarded DI. The decline is small among compliers. Displacement may be greater among applicants with the weakest labor-market prospects and the highest probability of receiving DI in the absence of YRA, but this group is underrepresented among compliers. Thus, the small effect here reflects the fact that the compliers' likely counterfactual state was not disability insurance.

YRA approval reduces employment and cumulative earnings during eligibility. Table 5 shows a decline in cumulative wage earnings by age 21, and Fig. 8 shows a decline in employment. At least some applicants would therefore have worked without YRA. The decline in cumulative earnings by age 21 amounts to 36% of the untreated complier mean. Early labor-market experience can be valuable, especially for young people with weak labor-market attachment (e.g. Heller, 2014; Hensvik et al., 2023; Knutsson and Tyrefors, 2024). The earnings decline likely reflects reductions in summer or part-time work, since few rejected applicants work full time. [17]

Instrumenting current YRA receipt rather than first-application approval yields a similar earnings decline, but the estimates become less precise as recipients leave the program and the first stage weakens (Fig. A.11).

Although the IV estimates are imprecise, they consistently indicate less work during eligibility. Because the benefit is temporary, these short-run responses help interpret its

[17] In Fig. A.15 I show that this is the kind of working spells the recipients most commonly have.

	(1) Cumulative schooling (years)	(2) Secondary degree	(3) Cumulative DI receipt (years)	(4) Cumulative wages (euros)
Panel A. Main IV estimates				
	0.005	**0.022**	**-0.149**	**-2,312***
	(0.162)	**(0.058)**	**(0.127)**	**(1,227)**
AR p-value	0.974	0.706	0.242	0.060
First-stage F	386	386	386	386
Panel B. IV with additional controls				
	0.046	0.041	-0.189	-1,844
	(0.157)	(0.056)	(0.118)	(1,208)
AR p-value	0.771	0.462	0.111	0.127
First-stage F	388	388	388	388
Panel C. OLS benchmark				
	-0.024	-0.076***	-0.248***	-2,890***
	(0.017)	(0.006)	(0.014)	(139)
Untreated complier mean	4.11	0.395	0.367	6,475
Sample mean	4.08	0.382	0.430	5,326
Individuals	30,420	30,420	30,420	30,420

Table 5: Main IV results on cumulative education and labor-market outcomes by age 21 (the end of benefit eligibility).

Notes: AR denotes the Anderson–Rubin test robust for weak instruments. Cumulative schooling is measured as years spent as a student after age 15. The sample size is the number of individuals observed through age 21. Standard errors are clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

long-run effects.

5.1.2 Long-run impacts

After eligibility ends, YRA approval reduces employment by 11.2 percentage points and raises NEET status by 11.7 percentage points. Table 6 presents the main labor-market results. I measure adult outcomes six to eleven years after application, roughly ages 21–28, when recipients are no longer subject to the program's education or rehabilitation condition. The regressions pool annual observations over this period.

This window captures adult labor-market outcomes in this sample. Although many Finns complete higher education between ages 25 and 30, relatively few applicants in my sample ever attend higher education. They are also on the margins of labor-market participation, so their earnings profiles are likely to be flatter than those of an average worker.

These effects amount to 28% and 26% of the untreated complier means, respectively. Both estimates are statistically significant. Adding extensive controls in panel B leaves the estimates nearly unchanged. The estimates remain statistically significant in most robustness specifications, with magnitudes close to the main estimates.

Annual earnings fall by €1,800, or 19% of the untreated complier mean. Applicants whose main activity is employment earn €20,028 per year on average, so the observed decline is smaller than the €2,243 implied by the employment change alone. This difference suggests that applicants induced not to work have lower earnings than the average employed applicant.

There is no measurable change in student status. The nearly equal decline in employment and rise in NEET status already imply little change in other activities. The point estimates for student status are close to zero, and the confidence intervals are relatively tight. Continuing education beyond secondary school is rare in this population, so applicants move mainly between employment and being out of employment, education, or training.

The OLS estimates are similar to, or slightly more negative than, the IV estimates. This pattern can reflect either selection into treatment or differences between the compliers and the full applicant sample. Table 6 shows that untreated compliers are more likely to be employed and have higher income than the full sample. I discuss these differences further in Section 5.1.5.

The employment and NEET effects persist after eligibility ends and are larger than during YRA receipt. Figure 8 presents the yearly IV estimates of the main labor-market outcomes. Effects on earnings follow a similar pattern (Panel C).

Disposable income increases during YRA receipt (panel D of Figure 8). This increase is smaller than the estimated increase in YRA payments because rejected applicants can receive other benefits. After the benefit ends, the effect on disposable income is negative (but less than the effect on earnings) or zero, as the welfare state replaces part of the lost earnings. Among applicants observed through age 27, the estimated increase in benefit receipt exceeds the estimated earnings loss (Section 6.2).

A more precise decomposition by benefit type is shown in Fig. A.4. Unemployment insurance and social-assistance payments increase the most, and both increases are statisti-

cally significant. This reflects the decline in labor-market participation. In contrast, there is no detectable change in pension or DI receipt. This null is also informative about the information and beliefs channel, a matter to which I return in the mechanisms section.

Applicants are on the margins of labor-market participation and may therefore respond strongly to YRA. The confidence intervals are wide but rule out large positive effects. The program therefore does not achieve its stated goal of increasing recipients' labor-market attachment.

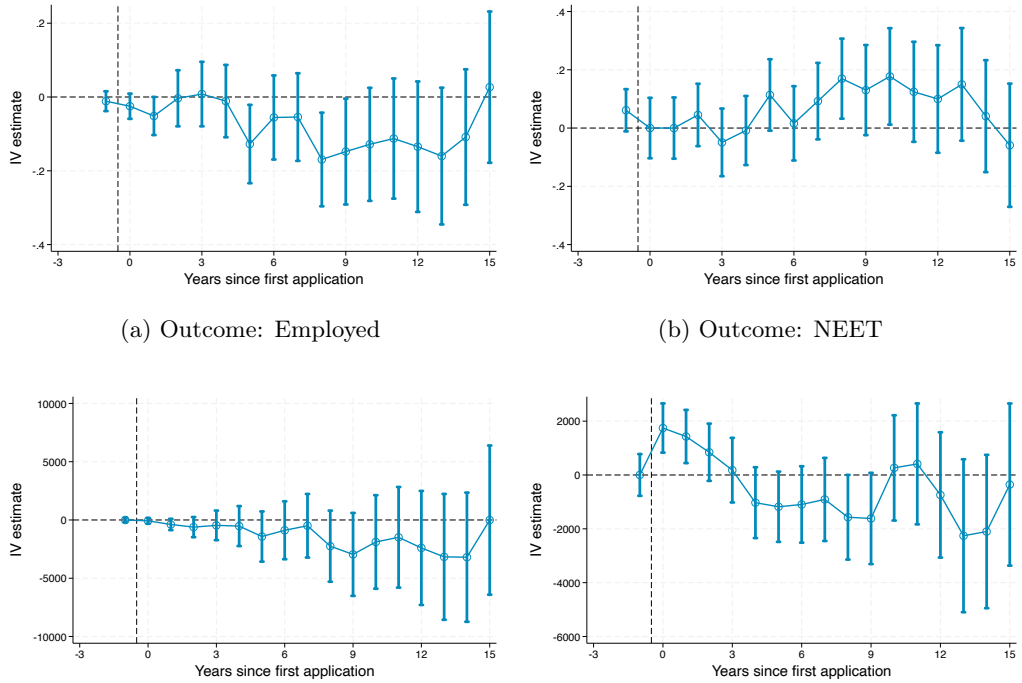
	(1) NEET	(2) Employed	(3) Wage income (€)	(4) Student
Panel A. Main IV estimates				
	0.117**	**-0.112****	**-1,756**	**-0.006**
	(0.054)	**(0.050)**	**(1,331)**	**(0.032)**
AR p-value	0.031	0.026	0.189	0.842
First-stage F	315	315	315	315
Panel B. IV with additional controls				
	0.100*	-0.098**	-1,455	-0.003
	(0.051)	(0.049)	(1,298)	(0.032)
AR p-value	0.052	0.043	0.263	0.927
First-stage F	319	319	319	319
Panel C. OLS benchmark				
	0.115***	-0.100***	-3,083***	-0.015***
	(0.003)	(0.003)	(77)	(0.002)
Untreated complier mean	0.454	0.399	9,177	0.147
Sample mean	0.567	0.295	6,802	0.138
Observations	102,595	102,595	102,595	102,595
Individuals	23,596	23,596	23,596	23,596

Table 6: Main IV results on labor-market outcomes in years 6–11 after the first application.

Notes: AR denotes the Anderson–Rubin test robust for weak instruments. Outcomes are measured during years 6–11 after the first application, which roughly corresponds to ages 21–28. Labor-market outcomes are available through 2022. Both OLS and IV include region × year × language fixed effects. IV with additional controls also includes disability category, indicators for depression, anxiety, F9, and other healthcare visits, urban residence, parental employment, parent ages, and ninth-grade GPA. The sample size refers to individual-year observations. Standard errors are clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5.1.3 Health and healthcare use

YRA approval does not detectably change hospital days, overnight mental-health admissions, or healthcare use during benefit eligibility (Table 7). Administrative healthcare data reflect both underlying health and care-seeking. I use the severity of the outcomes to distinguish between these dimensions. Hospitalizations and overnight mental-health admissions capture events that are serious enough that people are likely to seek care, so I treat them mainly as measures of underlying health. Intensive margin measures of visits and prescriptions also respond to resources and incentives to seek care. Healthcare co-payments are low or zero,



(a) Outcome: Employed

(b) Outcome: NEET

(c) Outcome: Wage earnings

(d) Outcome: Disposable income

Figure 8: Yearly IV estimates of the effects of YRA approval.

Notes: Each point comes from a separate EB-IV regression for the indicated event year. The outcomes are not recorded before age 16, so the pre-treatment estimates are pulled toward zero. Estimates are not scaled, so an estimate of 0.1 for a binary outcome means an increase of 10 percentage points.

especially for individuals under 18, but even small co-payments can reduce healthcare use (Haaga et al., 2024).

YRA could affect health and healthcare use in different ways. Additional income could reduce stress or increase use of healthcare, both of which would improve underlying health but have different effects on healthcare used. Reduced work may lower physical and mental strain, but it may also lead to a more passive lifestyle and more health issues, especially in the long run.

Columns 1 and 2 of Table 7 report effects on serious health outcomes over the years in which the benefit is most commonly received. I find no detectable effects on hospital days or overnight mental-health visits. The estimates are somewhat imprecise, but they provide no evidence of improvements in these measures of serious adverse health events.

I also find no evidence that the benefit substantially changes healthcare use. The IV coefficients are negative but not statistically significant, and small effects cannot be ruled out. This is consistent with the hypothesis that the individuals' healthcare use is not substantially constrained by money in the setting of universal health insurance with low out-of-pocket costs, especially for individuals under the age of 18.

Table A.1 presents health outcomes after the benefit ends. The estimates indicate a decline in mental-health visits, although this effect is not statistically significant. Overall, I find no evidence that the YRA affects health or healthcare use, though the confidence intervals include medium-sized effects. The estimates based on the conditional-ignorability approach in Section 5.2 are more precise.

	(1) Hospital days	(2) Overnight MH visits	(3) Any MH visit	(4) Any antidepressants
Panel A. Main IV estimates				
	0.326	**0.055**	**-0.028**	**-0.033**
	(0.904)	**(0.062)**	**(0.041)**	**(0.040)**
AR p-value	0.719	0.374	0.493	0.407
First-stage F	394	394	394	394
Panel B. IV with additional controls				
	0.257	0.051	-0.029	-0.030
	(0.884)	(0.062)	(0.038)	(0.039)
AR p-value	0.771	0.411	0.456	0.443
First-stage F	394	394	394	394
Panel C. OLS benchmark				
	0.339^{***}	0.020^{***}	0.029^{***}	-0.015^{***}
	(0.064)	(0.004)	(0.003)	(0.002)
Untreated complier mean	1.85	0.0797	0.368	0.244
Sample mean	1.94	0.118	0.411	0.24
Observations	152,014	152,014	152,014	152,014
Individuals	30,534	30,534	30,534	30,534

Table 7: Main IV results on health and healthcare use in years 0–5 after the first application (during benefit eligibility).

Notes: AR denotes the Anderson–Rubin test robust for weak instruments. Outcomes are measured years 0-5 after the first application, which corresponds to the time of benefit eligibility. Health outcomes are available until 2023. Both OLS and IV use the same, necessary controls. IV with additional controls uses controls for the disability category, indicators for any depression/anxiety/F9/other healthcare visits, urban area, parental employment, parent ages, and 9th grade GPA. Sample size refers to individual × year observations. Standard errors clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

5.1.4 Robustness of the IV estimates

I assess the sensitivity of the IV estimates to alternative estimators, instrument definitions, controls, and sample restrictions.

1. The allocation of examiners is explicitly documented to follow the procedure described in the above sections. Any sorting of applications somewhere against the general rules would be a threat to identification. To address potential sorting based on the application details, the main tables already include a large set of additional controls. In the appendix, I also residualize the instrument by disorder category and add year × month fixed effects. The results remain similar, reducing this concern.
2. A single region with exceptional mechanisms might drive the results. The estimates remain stable when I exclude each region in turn, assuring that no single region is driving the results (Fig. A.18).
3. JIVE may perform poorly with many covariates. I also report estimates from the Unbiased Jackknife IV estimator (UJIVE), following Kolesar (2013); Goldsmith-Pinkham et al. (2026). Because it requires constructing examiner leniency within the estimation sample, its first stage is noisier and its second-stage confidence intervals are wider. The estimates remain close to the baseline (Fig. A.16).
4. Monotonicity is more plausible for a binary instrument than for a continuous measure of leniency, as Mogstad and Torgovitsky (2024) argue. I therefore restrict the sample to applicants assigned to examiners in the top or bottom 25% of the leniency distribution and use assignment to the top group as the instrument. I repeat the exercise using the top and bottom 10%. Both specifications yield estimates close to the baseline (Fig. A.16).
5. Empirical Bayes shrinkage may drive the results. Using the conventional leave-one-out examiner-leniency instrument produces similar estimates, so Empirical Bayes shrinkage does not drive the results (Fig. A.16).
6. The results might be sensitive to the minimum number of cases per judge required to enter the sample. In Fig. A.17 I show that the estimates remain stable across these thresholds.
7. Changes in sample composition over time may affect the long-run estimates. In Table A.12 I show the long-run estimates remain similar when limiting the sample to only the balanced sample.

5.1.5 Compliers, MTE, and the limits of the IV

The IV estimates apply to applicants whose approval depends on examiner leniency. Table 8 compares compliers with the full applicant sample and with approved applicants. I estimate the shares of compliers with each characteristic following Bhuller et al. (2020). Compliers are more likely to be male and to have a developmental disorder, but less likely to have a mood disorder. Their diagnoses are less likely to lead to adult DI, and their parents have weaker labor-market attachment.

Treatment effects may vary with baseline labor-market prospects, but the local IV estimates are too imprecise to trace heterogeneity along this margin. Estimating these effects requires a different approach. One possible approach is the Marginal Treatment Effects (MTE) (Section 4.2.4). I estimate different versions of the MTE and report the

results in Section A.4. All specifications are consistent with the treatment being beneficial for those with the lowest unobserved resistance, but the magnitudes vary widely for different specification choices. These estimates should be taken as suggestive also because of the precision: the confidence intervals in the low-resistance end span the entire plausible interval of the binary outcome (not included for clarity). The estimated ATE is around -0.07 using the normal distribution assumption for the error terms, which is close to the ATE estimated using the selection-on-observables approach. The sensitivity and lack of precision confirm the inadequacy of the IV alone to answer the question about the effects on those with the lowest labor market prospects.

	Share among compliers (%)	Sample share (%)	Share among accepted (%)
Male	59.83	53.00	54.52
Mood disorder (F3)	7.28	13.86	12.48
Anxiety disorder (F4)	8.79	8.95	8.02
Developmental disorder (F8)	36.42	25.59	27.02
ADHD/Conduct disorder (F9)	12.48	14.12	16.63
Other mental-health disorder	6.00	6.12	4.11
Chromosomal/intellectual disorder	15.22	19.01	21.72
Other physical disorder	13.81	12.35	10.02
Urban	76.33	79.66	78.67
Mother employed	60.56	70.39	69.69
Father employed	66.57	72.29	72.10
High DI risk	12.46	18.91	21.39
Low DI risk	87.54	81.09	78.61

Table 8: Complier characteristics.

Notes: D_i indicates YRA approval, and Y_i denotes the characteristic listed in each row. Following Bhuller et al. (2020), I estimate the complier mean of Y_i from an IV regression of $Y_i(D_i - 1)$ on $D_i - 1$, using examiner leniency as the instrument. The table also reports the complier share implied by the first stage.

I therefore complement the IV with selection on observables. Under conditional ignorability and overlap, covariate adjustment yields more precise estimates for the broader applicant population and for groups underrepresented among compliers. Section 4.3 describes the approach.

5.2 Results using the Selection-on-Observables approach

The covariate-adjusted estimates vary with baseline labor-market prospects. On average, YRA approval reduces employment and increases NEET status. Among applicants with low labor-market prospects, however, it raises employment and reduces NEET status. I present the main results and then show that my selection-on-observables estimates track closely the IV estimates on the key outcomes.

5.2.1 Main results

On average, YRA approval worsens long-run labor-market outcomes, although the effects are smaller than the corresponding complier estimates. Table 9 displays the main long-run results for three groups. In row 1, I report the estimates for the aggregate applicant sample,

which I refer to as ATE. Approval lowers employment by 4.7 percentage points and raises NEET status by 4.6 percentage points. These effects are smaller than both the IV estimates and the DDML estimates for compliers. The covariate-adjusted estimates are also less negative than the OLS estimates in Table 6, consistent with negative selection into approval on observed characteristics. The estimates for long-run student status are close to zero and precisely estimated in all three groups.

By contrast, among applicants with low labor-market prospects, approval raises employment by 4.7 percentage points and reduces NEET status by 4.9 percentage points. The employment gain is about 29 percent of the group mean. Figure A.9 shows the same pattern across risk deciles. Employment effects are negative through most of the distribution but turn positive in the highest-risk deciles, while NEET effects move in the opposite direction. The highest-risk deciles also show the largest increases in years of education and the largest reductions in years on DI by age 21.

Among applicants with low labor-market prospects, selection on observables appears to attenuate the estimated effects. The negative Oster δ values indicate that adding controls moves the estimates away from zero. If selection on unobservables operates in the same direction, the reported effects are biased toward zero.

	(1) Employed	(2) NEET	(3) Empl. income (euros)	(4) Student
ATE	-0.047***	0.046***	-1632.184***	0.001
	(0.003)	(0.003)	(77.075)	(0.002)
Oster delta(0)	3.1	2.6	3.7	-0.043
Outcome mean	0.29	0.57	6802	0.14
N	102589	102589	102589	102589
Moderate labor market prospects	-0.074***	0.073***	-2306.439***	-0.000
	(0.004)	(0.004)	(95.259)	(0.003)
Oster delta(0)	5.2	5.3	6	2.7
Outcome mean	0.33	0.51	7951	0.16
N	79592	79592	79592	79592
Low labor market prospects	0.047***	-0.049***	679.445***	0.003
	(0.005)	(0.007)	(113.602)	(0.004)
Oster delta(0)	-41	-28	-12	-9.5
Outcome mean	0.16	0.77	2826	0.072
N	22997	22997	22997	22997
Controls	All	All	All	All

Table 9: Covariate-adjusted estimates obtained using double/debiased machine learning (DDML) (Chernozhukov et al., 2018).

Notes: All estimates in the table rely on the conditional ignorability assumption. Outcomes are measured in years 6–11 after the first application, a period that roughly corresponds to ages 21–28. Oster's $\delta(0)$ measures the strength of selection on unobservables relative to observables required to reduce the OLS estimate to zero (Oster, 2019). Negative values require unobserved selection to oppose observed selection. Estimation and SEs clustered at the individual level.

The earnings effects also differ across groups. The yearly estimates in Fig. 9 show lower earnings for the full sample but higher earnings among applicants with low labor-market prospects. For applicants with low labor-market prospects, the pooled estimate in Table 9 is an increase of about €679 per year, or roughly 24 percent of mean earnings.

Short-run effects also differ sharply by applicants' labor-market prospects (Table 10). Cumulative earnings fall among applicants with moderate labor-market prospects, while the estimate for applicants with low labor-market prospects is close to zero. Among applicants with low labor-market prospects, YRA approval also increases time in education by 0.70 years and raises secondary-school completion by age 21 by 6.9 percentage points. These results imply that many applicants in this group would otherwise have left education.

	(1) Cum. years schooling	(2) Secondary degree	(3) Cum. years on DI	(4) Cum. wages
ATE	0.186***	-0.042***	-0.514***	-2060***
	(0.017)	(0.006)	(0.016)	(145)
Oster $\delta(0)$	-3.6	6.5	-6.5	6.3
Sample mean	4.1	0.38	0.43	5320
N	30393	30382	30393	30393
Moderate labor market prospects	0.067***	-0.069***	-0.116***	-2557***
	(0.018)	(0.007)	(0.014)	(172)
Oster $\delta(0)$	-2.4	21	-5.7	10
Sample mean	4.2	0.41	0.24	6077
N	24669	24662	24669	24669
Low labor market prospects	0.701***	0.069***	-2.199***	59
	(0.051)	(0.013)	(0.063)	(239)
Oster $\delta(0)$	18	-148	6.6	2.9
Sample mean	3.3	0.25	1.2	2054
Controls	All	All	All	All
N	5724	5720	5724	5724
Age at obs.	20	20	20	20

Table 10: Short-run estimates obtained from adjusting for covariates using DDML.

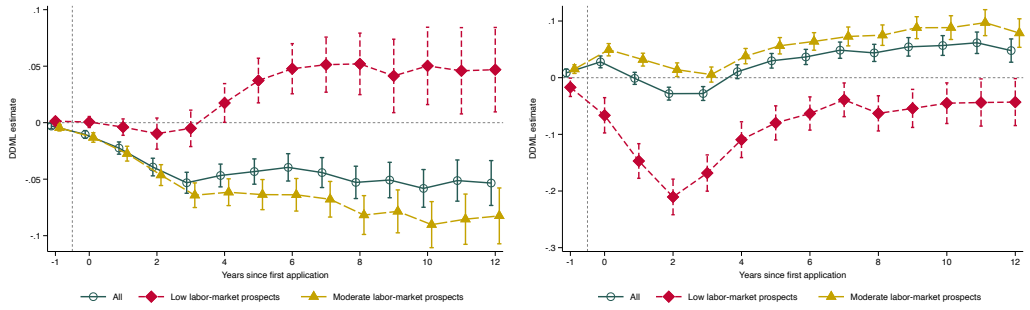
Notes: All estimates in the table rely on the conditional ignorability assumption. 'Cum. years of schooling' refers to years spent as a student after age 15. Estimation and SEs clustered at the individual level.

5.2.2 Validation

Two main checks support the conditional-ignorability design: the covariate-adjusted estimates are similar to the IV estimates for compliers, and the estimates change little across covariate sets.

Table 11 compares the IV and complier-weighted DDML estimates for the main labor-market outcomes. With all controls, DDML estimates imply a 9.0-percentage-point decline in employment and an 8.9-percentage-point increase in NEET status, compared with 11.1 and 11.7 percentage points, respectively, in the IV estimates. Both designs yield student-status estimates close to zero. Fig. A.22 presents year-by-year estimates with the two designs, which show very similar patterns. While I have limited power to detect very small differences, the similarity between the estimates supports the assumption that most of the relevant selection into treatment can be captured by the observables.

The IV comparison is informative only for the complier-weighted population. It does not directly support the estimates for applicants with low labor-market prospects, who



(a) Outcome: Employed

(b) Outcome: NEET

(c) Outcome: Wage earnings

(d) Outcome: Student

Figure 9: Main results using covariate adjustment.

Notes: Each point is a DDML estimate for the corresponding event year. All estimates in the table rely on the conditional ignorability assumption. Employment, NEET status, wage earnings, and student status are not recorded before age 16, so the figure does not provide a consistent set of pre-application estimates. Coefficients are reported in the outcomes' original units; for a binary outcome, a coefficient of 0.1 represents a 10-percentage-point increase. "All" refers to the entire overlap sample. "Low labor-market prospects" refers to applicants in the top fifth of predicted DI risk, and "moderate labor-market prospects" refers to the remaining applicants.

	(1)	(2)	(3)	(4)	(5)	(6)
Employed	-0.111**	-0.090***	-0.086***	-0.099***	-0.100***	-0.094***
	(0.050)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Oster $\delta(0)$		3.2				
Full-sample mean	0.29	0.29	0.29	0.29	0.29	0.29
NEET	0.117**	0.089***	0.092***	0.102***	0.109***	0.100***
	(0.054)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Oster $\delta(0)$		3.8				
Full-sample mean	0.57	0.57	0.57	0.57	0.57	0.57
Student	-0.007	-0.004	-0.003	0.001	-0.002	0.002
	(0.032)	(0.006)	(0.006)	(0.005)	(0.005)	(0.006)
Oster $\delta(0)$		-1.3				
Full-sample mean	0.14	0.14	0.14	0.14	0.14	0.14
Controls	IV controls	All	No parent	No applicant	No application	No healthcare
N	102586	22276	22276	22276	22276	22276

Table 11: Comparison of the IV estimates and covariate-adjusted DDML estimates on the complier-weighted sample 6–11 years after the first application (Chernozhukov et al., 2018).

Notes: The covariate sets are listed in Section 4.3. Columns 3–6 omit parental, applicant, application, and healthcare controls, respectively. The Oster $\delta(0)$ measures the strength of selection on unobservables needed to reduce the estimate to zero. For instance, $\delta(0) = 2$ means that selection on unobservables would need to be twice as strong as selection on observables. I calculate $\delta(0)$ using an OLS regression with all controls and set $R^2_{max} = 1.3\tilde{R}^2$. The measure is not informative for student status, for which the estimate is already close to zero, but I report it for completeness. Estimation and SEs clustered at the individual level.

are underrepresented among compliers. For this group, the case for the design rests on propensity-score overlap and stability across covariate sets.

The estimates also change little across covariate sets. Omitting application controls produces the largest change in the employment and NEET estimates: the employment effect changes from -9.0 to -10.0 percentage points, and the NEET effect changes from 8.9 to 10.9 percentage points. I also report the Oster δ coefficients. For employment and NEET status in the complier-weighted subsample (Table 11), selection on unobservables would need to be 3.2 and 3.8 times as strong as selection on observables, respectively, to reduce the estimates to zero.

5.2.3 Other outcomes

YRA approval has no clear effects on crime or overnight mental-health visits in either low-prospect or other applicant group (Fig. A.8). The selection-on-observables approach provides more precise estimates for these outcomes. These results are broadly in line with the IV estimates. There is a clear gradient of responses by the DI index in disability insurance and probability of being married, in the direction of the main outcomes.

5.3 Effects on parents and siblings

Previous research in related settings has shown that welfare benefits resembling the Youth Rehabilitation Allowance can have effects beyond the original recipient (Dahl et al., 2014;

Deshpande, 2016b; Autor et al., 2019).

I first take a look at how the parents respond to their child being approved to the benefit and the rise in the household income as a result. I use a differences-in-differences design, comparing the approved and rejected applicants before and after the decision. I find that employment earnings decline for both mothers and fathers, but the declines are small and not significant for fathers (Fig. A.7). The estimates are precise and decreases larger than 2.7% and 1.8 of the approved mean earnings prior to treatment for mothers and fathers, respectively [18]. Effects as large as in Deshpande (2016b) can be rejected. There are no detectable extensive margin responses and the effects in parent disability insurance receipt are null as well.

This result should not be interpreted as a pure income effect of a transfer, since it might also change things about the child. These results are consistent with the assumption that the child and the parents have different budgets, so the parents respond only a little.

I then turn to siblings and estimate the effects of sibling being admitted to the Youth Rehabilitation Allowance on other siblings applying or receiving Youth Rehabilitation Allowance. This tests whether siblings potentially transmit beliefs and information about the benefits to each other. I find increases in both applications and YRA received. These findings are in line with the previous research (Dahl et al., 2014; Deshpande, 2016a,b).

6 Discussion, mechanisms, and welfare effects

6.1 Mechanisms

The benefit reduces later employment among marginal applicants, while the evidence suggests gains for those with the lowest prospects. In this section, I examine whether changes in schooling and early work can explain this difference, then consider alternative mechanisms.

The effects on the labor market are still visible a decade after the payments have stopped. This rules out the usual explanation in the adult disability insurance literature, where the negative effects follow from the income and substitution effects of a benefit currently being paid (Bound, 1991; Maestas et al., 2013; Autor et al., 2019). The YRA is paid mostly for four years and ends around age 20. The explanations for the effects observed in previous sections must therefore be something that changed during the benefit recipiency.

The heterogeneity in Section 5.2 suggests a possible direction. Every accepted applicant receives the same transfer, and the screening and the conditions are the same. The applicants differ in the counterfactuals: what they would have done without the benefit. For the compliers of the IV design, who have moderate labor market prospects, the benefit reduced work: cumulative wage earnings by age 21 fall by 36 % of the untreated complier mean, while not shifting participation in education by much (Table 5). These individuals were not displaced from the Disability Insurance, as intended by the policymakers.[19]

[18] From Table 1. Mothers: $\frac{-536}{19,697}$ and fathers $\frac{-506}{26,742}$.

[19] Interestingly, in informal discussions I had with the social workers who are in touch with the applicants, this also came up. Many social workers did not think the applicants were even near the margin to qualify for disability insurance.

For applicants with the lowest labor market prospects, the benefit replaces disability insurance, average earnings during eligibility respond little, and the long-run employment effects are positive (Table 9 and Fig. 9). Lost work experience has also been proposed as an explanation for the negative effects of other temporary cash transfers (Barr et al., 2021; Price and Song, 2025; Britton et al., 2025). Filippucci (2025) finds that active labor-market measures offset the negative effects of a cash transfer by changing recipients' activity. Below, I use a standard human-capital model (Becker, 1964; Mincer, 1974) to relate changes in schooling and early work to adult employment.

Previous literature on cash transfers also offers possible other explanations. One is the formation of 'Welfare cultures', so changes in beliefs, information or stigma related to welfare use (Moffitt, 1983; Lindbeck et al., 2003; Dahl et al., 2014). A second one is that cash can have negative effects on health (Dobkin and Puller, 2007; Evans and Moore, 2011; Black et al., 2024), possibly through consumption of harmful goods. Below, I will also consider the evidence for these theories.

6.1.1 Employment: education and early work

The YRA can change what applicants do between ages 16 and 20. Approval combines a cash transfer with a requirement to attend education or rehabilitation. It may keep applicants in school instead of receiving disability benefits, or reduce the work they would have done alongside their studies. Schooling and work build different skills, so the effect on later employment depends on which activity the benefit replaces.

I summarize these ideas with a simple human-capital model. At the end of compulsory school, applicant i has an initial stock of academic human capital, proxied by ninth-grade GPA, G_i. Let $d \in \{0, 1\}$ denote YRA approval. Adult human capital under approval state d is

$$H_i^A(d) = h(G_i) + \theta_S S_i(d) + g(X_i(d)), \qquad \theta_S \geq 0, \quad g' > 0, \quad g'' < 0. \tag{7}$$

where S_i is subsequent time in education, and X_i is productive early labor-market exposure. The coefficient θ_S denotes the effect of schooling on the human-capital index. The function $g(X_i)$ captures the standard Mincerian idea that workers acquire skills through experience at a declining rate (Mincer, 1958, 1974; Dustmann and Meghir, 2005).

The paper's main adult outcome is employment.[20] I therefore add a participation decision. Applicant i is employed as an adult under treatment state d when the offered wage exceeds the minimum offer needed to work:

$$E_i^A(d) = \mathbf{1}\{\alpha + \lambda H_i^A(d) + u_i \geq R_i\}, \qquad \lambda > 0. \tag{8}$$

Here $\log w_i^*(d) = \alpha + \lambda H_i^A(d) + u_i$ is the log wage offer, and u_i is a wage-offer shock. The reservation index R_i summarizes health, the cost of working, and available benefits in adulthood.[21] Schooling and early work raise the wage offer and can therefore move an

[20] I do not observe hourly wages or hours worked. A wage regression among employed applicants would also describe a selected group.

[21] I hold R_i fixed across treatment states here. Approval could also affect this threshold through changes in health or attitudes toward work, which I discuss below. This could be modeled with R being a function of d.

applicant into employment.

The YRA affects this decision by changing the human-capital inputs accumulated before adulthood. Its effect on applicant i's human capital is

$$\Delta H_i \equiv H_i^A(1) - H_i^A(0) = \theta_S \Delta S_i + \Delta g_i. \tag{9}$$

[22] Here $\Delta S_i = S_i(1) - S_i(0)$, and $\Delta g_i = g(X_i(1)) - g(X_i(0))$ denote the corresponding changes in the human-capital inputs.

Because a higher wage offer raises employment, a first-order approximation maps the average change in human capital into the average change in employment:

$$\Delta p \approx \underbrace{a\theta_S \Delta \bar{S}}_{\text{education channel}} + \underbrace{a\Delta \bar{g}}_{\text{early-work channel}}. \tag{10}$$

Here Δp is the average change in the adult employment probability, $\Delta \bar{S}$ and $\Delta \bar{g}$ are the average changes in schooling and early-work human capital, and $a \geq 0$ scales changes in human capital into changes in employment.[23]

The model predicts lower adult employment when approval reduces early work without increasing education. If approval moves applicants from disability insurance into education without displacing work, adult employment should rise. When education increases and work falls, the net effect depends on the size of each change and the returns to the two inputs. Studying and working can occur in the same year, so changes in the two inputs need not offset each other one-for-one.

Empirical implementation

I combine the effects of YRA approval on schooling and early work with descriptive employment returns to these inputs. This is a similar exercise to (Heckman et al., 2013) and is on a broad level similar to controlling for the mediators in a regression.

Schooling is years recorded as a student through age 20; cumulative earnings through age 20 measure early work attachment.[24] As Ashworth et al. (2021) emphasize, using realized experience is preferable to potential experience (e.g. age − years in education). I use earnings as a proxy because hours worked and short employment spells are not reliably observed.

I estimate the descriptive returns separately by applicant group, pooling approved and rejected applicants. To allow a flexible relationship between early work and adult employment, I divide cumulative earnings into five categories, including a separate category

[22] Ninth-grade GPA does not appear as a treatment-induced change because it is measured before the application. It can still determine the initial stock of human capital.

[23] Formally, $\Delta p = \mathbb{E}[p_i(1) - p_i(0)]$, $\Delta \bar{S} = \mathbb{E}[\Delta S_i]$, and $\Delta \bar{g} = \mathbb{E}[\Delta g_i]$. If m_i is the slope of applicant i's employment probability at the baseline threshold, then $a = \lambda \mathbb{E}[m_i]$. The approximation requires this probability to be differentiable and $\mathrm{Cov}(m_i, \Delta H_i) = 0$.

[24] I estimate the input effects in the long-run person-year sample (years 6–11), repeating the age-20 stocks, rather than using one observation per person as in the short-run tables.

for zero earnings:

$$E_i^A = \beta_{0,k} + f_k(G_i) + \beta_{S,k} S_i + \sum_{b=0}^{3} \beta_{b,k} D_{bi} + C_i' \gamma_k + \varepsilon_i, \tag{11}$$

where D_{bi} indicates earnings category b, with the highest category omitted. The term $f_k(G_i)$ includes ninth-grade GPA and its missing indicator. The controls C_i include calendar-year effects and, for the lowest-prospect group, age indicators. For the IV projection, I estimate the returns in the sample with high complier weights. I interpret the coefficients as descriptive associations.

For each group k, let $\Delta\bar{S}_k$ denote the effect of approval on schooling and $\Delta p_{b,k}$ its effect on the probability of belonging to earnings category b. The implied employment effect is

$$\widehat{\Delta p}_k^{\,\text{proj}} \equiv \underbrace{\widehat{\beta}_{S,k}\widehat{\Delta\bar{S}}_k}_{\text{schooling contribution}} + \underbrace{\sum_{b=0}^{3} \widehat{\beta}_{b,k}\widehat{\Delta p}_{b,k}}_{\text{early-work contribution}} . \tag{12}$$

I compare this projection with the estimated employment effect for each group. Because the returns are descriptive, the contributions do not identify causal mediation.

	Schooling	Early work	Implied total	Actual effect	Share reproduced
Marginal applicants (IV)	−0.28	−3.62	−3.90	−11.20 (5.00)	35%
Other applicants (DDML)	0.06	−5.77	−5.71	−7.40 (0.40)	77%
Lowest prospects (DDML)	1.52	1.68	3.20	4.70 (0.50)	68%

Table 12: Descriptive accounting for adult employment effects.

Notes: Contributions combine the effects of approval on age-20 inputs with group-specific descriptive returns estimated among approved and rejected applicants. Input effects use IV for marginal applicants and DDML for the other groups. All effects are in percentage points; shares compare implied and actual effects. Parentheses report clustered standard errors for the actual effects (Tables 6 and 9).

The projections in Table 12 have the same sign as the employment effects for all three groups. For marginal applicants, the implied decline is 3.9 percentage points, about one-third of the estimated 11.2-point decline. Reduced early work contributes 3.6 points, while schooling contributes little. For other applicants, early work accounts for almost the entire projected decline of 5.7 points, about three-quarters of their estimated 7.4-point decline. The projection for applicants with the lowest prospects is a 3.2-point increase, about two-thirds of the estimated 4.7-point gain. Schooling contributes 1.5 points and early work 1.7 points. Mean early earnings change little in this group, but approval reduces the share with zero earnings. The descriptive returns assign a positive employment contribution to this change in work attachment.

Previous studies also find that employment while in school predicts better subsequent labor-market outcomes, although the estimates are sensitive to how selection into youth

employment is addressed (Ruhm, 1997; Hotz et al., 2002). Direct causal estimates of returns to early work attachment are scarce. Studies of early job matches, summer-job programs, and subsidized employment find returns that vary across settings (Hensvik et al., 2023; Knutsson and Tyrefors, 2024; Heller, 2014; Gelber et al., 2016; Le Barbanchon et al., 2020).[25] Bruhn et al. (2025) use quasi-random variation in early occupational experience outside subsidized employment and find large and persistent effects of initial exposure to a profession.

A change in adult health, work costs, or benefit options would operate through R_i rather than H_i^A. If approval changes this threshold by ΔR_i, the employment index changes by $\lambda \Delta H_i - \Delta R_i$. I examine those channels separately using the adult health and benefit outcomes. The exercise here asks how much of the employment pattern is consistent with the changes in education and early work documented during the benefit period.

I also conduct a descriptive mediation analysis, where I include the accumulated earnings by age 20 in the equation where the outcome is adult employment. Table A.6 presents the results. The DDML estimates for the whole sample and for the complier-weighted population change substantially, which would be consistent with a large role for the mediator. The IV estimates change less when the cumulative wage earnings are included. These numbers should be interpreted carefully and as descriptive: the mediator is not randomized and in the IV specification the adjusted treatment coefficient need not retain the original complier-LATE interpretation.

6.1.2 Information, beliefs or other forms of welfare dependency

Another possibility is that receiving the YRA changes how recipients interact with the social insurance system. It may teach them how to apply for related benefits or reduce the stigma of claiming them (Goffman, 1963; Moffitt, 1983). It could also change attitudes toward work and self-sufficiency or medicalization of "normal" life situations (like poor focus in school or negative emotions stemming from the current environment). While this type of story is common in policy debates, there is little credible empirical evidence isolating this type of channels, making it harder to formulate a testable hypothesis about how it would operate and for which groups it would be especially prevalent.

If the YRA teaches recipients how to navigate health-conditioned benefits or reduces the stigma of claiming them, adult disability insurance is the natural outcome to examine. Adult DI resembles the YRA in its health conditions and application requirements, although its assessment is separate and more demanding. I observe no increase in DI recipiency after the benefit ends in the IV analysis (Fig. A.4). Using the selection-on-observables approach, a small increase is detected for the overall group, but not enough to explain the employment effects (Fig. A.8)[26]. Looking at another health-based benefit, the more short-term sickness benefit, I also find null effects (Fig. A.4). Unemployment insurance and social assistance, the general benefits available to anyone out of work, increase instead, which follows from the decline in employment rather than from anything specific to the YRA.

[25] The YRA displaces work in the ordinary labor market. Connections formed in these jobs may differ from those created by subsidized employment programs.

[26] And much smaller than in Dahl et al. (2014) where the treatment is parent being approved to a benefit (there DI), not the focal child earlier in life, like in my setting.

Based on the previous literature, we would expect the opposite pattern. Dahl et al. (2014) find that a parent receiving disability insurance has a large effect on their children eventually receiving it as well, an effect hypothesized to work through information about the application process and beliefs about success in the applications, and one that operates through the same benefit and even the same diagnosis codes. For the informational channel there is also evidence from other settings (Finkelstein and Notowidigdo, 2019; Anders and Rafkin, 2024).[27]

The administrative outcomes leave two limitations. The IV confidence intervals include meaningful increases in DI recipiency, and there could be an increase in DI applications, which I do not observe in my data. Beyond that, these channels are hard to test without detailed survey data on information, attitudes, and beliefs, which I do not have. The existing qualitative work does not support this kind of changes[28], but these hypotheses remain untestable.

6.1.3 Changes in health and consumption of harmful goods

The recipients could also spend the cash on harmful or addictive goods. Stories of this kind are documented by Evans and Moore (2011) and Chorniy et al. (2025), and suggested as a reason for the negative health effects of disability insurance by Black et al. (2024). My data allow me to test this directly. Using crime data, I test whether the accepted are more likely to be convicted of or reported to the police for drug-related crimes. Using healthcare data, I test whether they are more likely to be diagnosed with a substance abuse disorder in specialized or primary care; these data do not include stays in drug detox or rehabilitation institutions, but a diagnosis from healthcare typically precedes such a stay. In either data source and with either identification strategy, I find no increase in substance abuse (Fig. A.3). If anything, there is a decrease in the first years of benefit recipiency.

A possibility in a similar vein is that the benefit allows the recipients to get used to a level of consumption that exceeds their post-benefit ability to fund this consumption. They would then accumulate debt after the benefit ends, which could lead to financial distress and garnishment of wages, and inhibit the later labor market outcomes. I find no evidence of increased wage garnishment in Fig. A.3. If anything, there is reduction during the benefit receipt, although not statistically significant.

6.2 Welfare effects

The MVPF framework of Hendren and Sprung-Keyser (2020) is an approach to accounting for the benefits and costs of a public policy. The MVPF is the willingness to pay of recipients and affected third parties per euro of net cost to the government. In my setting I

[27] A mechanism like this should not be understood to necessarily be a bad outcome: the welfare impact would depend on many things, including how the channel affects targeting of the social insurance benefits. In many informational experiments, the recipients induced to take up a benefit are more likely to be eligible than the average recipient, or the ones with the highest need face the highest barriers (Finkelstein and Notowidigdo, 2019; Castell et al., 2025).

[28] Evidence from Basic Income experiments generally does not suggest large changes in attitudes (Blomberg et al., 2021; Broockman et al., 2024), and previous qualitative research on YRA recipients does not document attitudes of this kind (Salminen, 2019).

can observe, among other things, the use of healthcare and criminal activity, allowing for a detailed accounting of the expenses. I also observe all benefits paid to the individuals.

$$\text{MVPF} = \frac{\text{WTP}}{\text{Net government cost}} = \frac{\sum_i \text{WTP}_i}{M - FE}, \tag{13}$$

where M is the mechanical cost of the transfer, holding behaviour fixed, and FE is the fiscal externality, the change in the government budget caused by the behavioural responses to the policy:

$$FE = \underbrace{\Delta\tau}_{\text{taxes}} - \underbrace{\Delta B}_{\text{other transfers}} - \underbrace{\Delta H}_{\text{healthcare}} - \underbrace{\Delta CJ}_{\text{criminal justice costs}} - \underbrace{\Delta E}_{\text{Costs of providing education}}. \tag{14}$$

Under this sign convention, behavioural responses that save the government money, such as higher tax revenue or lower spending on healthcare and the criminal justice system, reduce the net cost and raise the MVPF.

The largest category of government expenses is the social insurance transfers paid to the individuals. In the MVPF, this amount features on both sides of the ratio: recipients value a euro of transfer at approximately a euro, so the transfer enters the numerator at its cost and the mechanical cost M in the denominator, and everything I estimate is a deviation from that benchmark. I discount all amounts at an annual rate of 3% (as Hendren and Sprung-Keyser (2020)) and present them in 2023 euros.

I exclude earnings during benefit recipiency from the calculation because I assume these losses reflect individuals choosing to work less due to the income effect of the benefit. I instead include earnings losses after the benefit ends in willingness to pay, since, in my judgment, they reflect at least partly lower productivity. While I prefer this interpretation, the long-run earnings response could instead reflect a choice to work less. The alternative calculation in the table excludes post-program earnings changes and all changes in other benefit receipts from willingness to pay, leaving YRA payments plus the benefits to crime victims. The net government cost is unchanged: changes in taxes and other transfers still enter the denominator. For compliers, this calculation treats the additional benefit receipts as part of the same labor-supply response. As shown in the last row of Table 13, this choice has a substantial effect on the MVPF.

I include only the costs of hospital and specialized healthcare as well as the drug costs to the government in the cost estimates, since the costs of primary care are not available. I do not changes in health on the WTP side of the equation, since it would require mapping the healthcare data into quality-adjusted life years or similar measures, and estimates indicate small effects, if any. Finally, the calculations assume zero effects beyond age 27, an age chosen because the later ages rely on estimates that only represent a few cohorts, which might not be representative of the whole sample. For a more detailed discussion of the choices and associated caveats, I refer the reader to Section A.6.5.

With my preferred choices, the benefit has a marginal value of public funds of 0.8 (Table 13). A pure transfer that induced no behavioral response would have an MVPF of one, so the YRA is not an efficient way of moving resources to this group: to justify it, we would need the recipients to have a 1.2 times larger welfare weight than the general

population.

Instead, on the individuals with the lowest prospects, the government costs are negative. This means that the policy pays for itself and the MVPF, as defined by Hendren and Sprung-Keyser (2020), is defined to be infinite. The majority of this comes from the increase in tax revenue from these individuals. While there is also a large change in other benefits paid, this enters both sides of the MVPF formula. This number is estimated using the DDML procedure, and it inherits the strong identification assumptions highlighted in earlier sections.

	Euros per recipient	
	IV (compliers)	Low prospects
Panel A. Willingness to pay		
YRA payments	16559	19894
Other benefits	3991	-21771
Change in earnings	-8731	1756
Benefits to crime victims	2365	953
Total	14184	832
Panel B. Net cost to the government		
YRA payments	16559	19894
Other benefits	3991	-21771
Reduction in tax revenue	1765	-1359
Healthcare costs	-3922	1746
Criminal justice costs	-694	752
Total	17700	-738
MVPF	**0.80**	**-1.13** (∞)
MVPF, without earnings loss	**1.07**	**-28.25** (∞)
Benefit–cost ratio (per euro of YRA payments)	**0.79**	**1.08**

Table 13: MVPF calculation, using the sample of individuals observed until age 27.

Notes: IV and DDML estimates. Healthcare costs include hospital/specialized healthcare visits and drug expenses. The row 'without earnings loss' excludes post-program earnings changes and other benefit changes from willingness to pay, retaining YRA payments and benefits to crime victims; the net government cost is unchanged. Positive entries in panel A increase willingness to pay, while positive entries in panel B increase government costs. Benefits to crime victims are reductions in victimization harm. All values are in 2023 euros and discounted at an annual rate of 3% (as in Hendren and Sprung-Keyser (2020)). Values may not sum because of rounding.

7 Conclusion

I study whether Finland's Youth Rehabilitation Allowance, a temporary cash transfer tied to education or rehabilitation, improves the adult outcomes of adolescents with health impairments. The program aims to keep recipients in education rather than on disability benefits. It's details allow me to test interesting predictions about cash transfers in contexts not widely studied prior to this. I show that the employment effects differ sharply across applicants.

I use examiner leniency as an instrument for YRA approval to estimate effects on marginal applicants. Approval reduces employment and increases benefit receipt after the benefit ends in this population. The estimates from this design are local to marginal applicants (IV compliers): applicants whose approval depends on the assigned examiner.

These marginal applicants constitute a substantial share of the applicant population. The adverse effects indicate that the program does not achieve the goals of the policymakers.

Applicants with low labor-market prospects are underrepresented among marginal applicants. I study this group by comparing approved and rejected applicants after adjusting for observed differences. The selection-on-observables estimates closely replicate the IV estimates in a sample resembling marginal applicants in observed characteristics. The agreement between estimates, together with various placebo tests, supports a causal interpretation of the selection-on-observables estimates. The estimates suggest that approval improves adult employment among applicants with low labor-market prospects. The short-run responses help explain why the long-run effects differ from those for marginal applicants. YRA moves these applicants from disability benefits to education and does not reduce average earnings during the benefit. The gains for applicants with low labor-market prospects show that public policy can improve their employment outcomes.

Improving the labor market outcomes of individuals with health impairments requires understanding of what drives their labor market outcomes. During the benefit period, the program studied here affects both time spent in education and attachment to the labor market. I show that a simple model can explain a substantial share of the policy's effects on later employment. While the model is deliberately simple, this result suggests that more research on the determinants of early labor market success for disadvantaged groups could help understand the barriers these individuals face as well as devise solutions.

The two empirical approaches illustrate how quasi-experimental or experimental and selection-on-observables designs can complement each other (see also Chetty et al., 2026). The selection-on-observables approach extends the analysis to applicants underrepresented among marginal applicants, subject to the stronger identifying assumptions required by the selection-on-observables approach. Although my setting is particularly favorable because of its credible control pool and rich observed covariates, the results suggest that combining these approaches may also be useful in other settings.

The results in this paper should be of interest in settings beyond mine. The rapid increase in the share of youth with health impairments is a wide phenomenon in high-income countries. While the Youth Rehabilitation Allowance in itself is a novel policy for economics literature, the adult disability programs and incentive problems surrounding them are a common feature of social insurance systems throughout high-income countries (Maestas et al., 2013; Dahl et al., 2014; Borghans et al., 2014; Autor et al., 2019). Thus similar tradeoffs between income support and incentives to invest in human capital arise in other settings as well. The mechanisms proposed here are a way to reconcile some disappointing results on conditional cash transfers in high-income countries (Britton et al., 2025) to more positive ones (Filippucci, 2025; Parker and Todd, 2017).

My estimates imply an MVPF below 1 for marginal applicants and net government savings for applicants with low labor-market prospects. The group with low labor-market prospects is defined using characteristics observed by policymakers, namely the applicant's diagnosis. Together, these facts suggest a realistic scope for improving targeting, although I do not take a stance on second-order consequences of such reforms. A broader lesson from public economics is that benefit targeting is important and observed characteristics can be used in this (Mirrlees, 1971; Akerlof, 1978). In the spirit of this work, in my setting

the applicants' diagnoses are informative about their likely activities in the absence of the benefit, and therefore about the long run effects as well. This is of importance when designing policies to support disadvantaged groups.

References

Agan, A., Doleac, J. L., and Harvey, A. (2023). Misdemeanor prosecution. *The Quarterly Journal of Economics*, 138(3):1453–1505. 14

Ahrens, A., Hansen, C. B., Schaffer, M. E., and Wiemann, T. (2024). ddml: Double/debiased machine learning in Stata. *The Stata Journal*, 24(1):3–45. 22

Aizer, A., Cho, S., Eli, S., and Lleras-Muney, A. (2024). The impact of cash transfers to poor mothers on family structure and maternal well-being. *American Economic Journal: Applied Economics*, 16(2):492–529. 25

Akerlof, G. A. (1978). The economics of" tagging" as applied to the optimal income tax, welfare programs, and manpower planning. *The American economic review*, 68(1):8–19. 46

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184. 21, 24

Anders, J. and Rafkin, C. (2024). The welfare effects of eligibility expansions: Theory and evidence from snap. *American Economic Journal: Economic Policy*. Forthcoming. 43

Andresen, M. E. (2018). Exploring marginal treatment effects: Flexible estimation using stata. *The Stata Journal*, 18(1):118–158. 78

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753. 16

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288. 23

Ashworth, J., Hotz, V. J., Maurel, A., and Ransom, T. (2021). Changes across cohorts in wage returns to schooling and early work experiences. *Journal of labor economics*, 39(4):931–964. 40

Attanasio, O. P., Meghir, C., and Santiago, A. (2012). Education choices in mexico: using a structural model and a randomized experiment to evaluate progresa. *The Review of Economic Studies*, 79(1):37–66. 4

Autor, D., Kostøl, A., Mogstad, M., and Setzler, B. (2019). Disability benefits, consumption insurance, and household labor supply. *American Economic Review*, 109(7):2613–2654. 1, 3, 13, 38, 46

Bald, A., Chyn, E., Hastings, J., and Machelett, M. (2022). The causal impact of removing children from abusive and neglectful homes. *Journal of Political Economy*, 130(7):1919–1962. 13

Barr, A., Kawano, L., Sacerdote, B., Skimmyhorn, W., and Stevens, M. (2021). You can't handle the truth: The effects of the post-9/11 gi bill on higher education and earnings. Technical report, National Bureau of Economic Research. 3, 39

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, by Gary S. Becker...*, volume 1. London. 39

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of political economy*, 128(4):1269–1324. 18, 20, 24, 32, 33, 73, 75

Black, B., French, E., McCauley, J., and Song, J. (2024). The effect of disability insurance receipt on mortality. *Journal of Public Economics*, 229:105033. 3, 4, 39, 43

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2026). When is tsls actually late? *The Review of Economic Studies*. 14

Blomberg, H., Kroll, C., and Tarkiainen, L. (2021). Life on basic income: Interview accounts by basic income experiment participants on the effects of the experiment. In Kangas, O., Jauhiainen, S., Simanainen, M., and Ylikanno, M., editors, *Experimenting with Unconditional Basic Income: Lessons from the Finnish BI Experiment 2017–2018*, chapter 12, pages 150–168. Edward Elgar Publishing, Cheltenham. 43

Borgan Reiling, R., Vea Salvanes, K., and Jorde Sandsør, A. M. (2024). Back to work? the impact of education as vocational rehabilitation for work-impaired youth. *Journal of Education and Work*, 37(5-6):433–457. 4

Borghans, L., Gielen, A. C., and Luttmer, E. F. (2014). Social support substitution and the earnings rebound: Evidence from a regression discontinuity in disability insurance reform. *American economic Journal: economic policy*, 6(4):34–70. 3, 46

Bound, J. (1991). The health and earnings of rejected disability insurance applicants: reply. *The American Economic Review*, 81(5):1427–1434. 3, 38

Britton, J., Villa, C., Ridpath, N., and Waltmann, B. (2025). Paying disadvantaged teenagers to stay in school. *IFS Working Paper 25/06*. 3, 26, 39, 46

Broockman, D. E., Rhodes, E., Bartik, A. W., Dotson, K., Miller, S., Krause, P. K., and Vivalt, E. (2024). The causal effects of income on political attitudes and behavior: A randomized field experiment. NBER Working Paper 33214, National Bureau of Economic Research. 43

Bruhn, J., Fabian, J., Gallagher, L., Gudgeon, M., Isen, A., and Phipps, A. R. (2025). Path dependence in the labor market: The long-run effects of early career occupational experience. Technical report, National Bureau of Economic Research. 4, 42

Card, D., Kluve, J., and 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. 4

Carneiro, P., Heckman, J. J., and Vytlacil, E. (2010). Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica*, 78(1):377–394. 20

Castell, L., Gurgand, M., Imbert, C., and Tochev, T. (2025). Take-up of social benefits: experimental evidence from france. *American Economic Journal: Economic Policy*, 17(4):1–29. 43

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. 2, 22, 34, 37

Chernozhukov, V., Cinelli, C., Newey, W. K., Sharma, A., and Syrgkanis, V. (2026). Long story short: Omitted variable bias in causal machine learning. *The Review of Economics and Statistics*, pages 1–45. 81

Chetty, R., Fogel, J., Katz, L. F., Noray, K., Porter, S., and Reisinger, J. (2026). Experimentally validated digital twins: Evaluating workforce training programs in the age of big data. *Working Paper*. 22, 46

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review*, 104(9):2593–2632. 15, 87

Chorniy, A., Finkelstein, A., and Notowidigdo, M. (2025). Paternalistic social assistance: evidence and implications from cash vs. in-kind transfers. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2025-145). 2, 43

Chyn, E., Frandsen, B., and Leslie, E. C. (2024). Examiner and judge designs in economics: A practitioner's guide. Technical report, National Bureau of Economic Research. 14, 16, 18

Dahl, G. B. and Gielen, A. C. (2021). Intergenerational spillovers in disability insurance. *American Economic Journal: Applied Economics*, 13(2):116–150. 2

Dahl, G. B., Kostøl, A. R., and Mogstad, M. (2014). Family welfare cultures. *The Quarterly Journal of Economics*, 129(4):1711–1752. 2, 37, 38, 39, 42, 43, 46

Dearden, L., Emmerson, C., Frayne, C., and Meghir, C. (2009). Conditional cash transfers and school dropout rates. *Journal of Human Resources*, 44(4):827–857. 3

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161. 21

Denning, J. T., Marx, B. M., and Turner, L. J. (2019). Propelled: The effects of grants on graduation, earnings, and welfare. *American Economic Journal: Applied Economics*, 11(3):193–224. 3

Deshpande, M. (2016a). Does welfare inhibit success? the long-term effects of removing low-income youth from the disability rolls. *American Economic Review*, 106(11):3300–3330. 1, 3, 38

Deshpande, M. (2016b). The effect of disability payments on household earnings and income: Evidence from the ssi children's program. *Review of Economics and Statistics*, 98(4):638–654. 38

Deshpande, M., Gross, T., and Su, Y. (2021). Disability and distress: The effect of disability programs on financial outcomes. *American Economic Journal: Applied Economics*, 13(2):151–178. 4

Deshpande, M. and Mueller-Smith, M. (2022). Does welfare prevent crime? the criminal justice outcomes of youth removed from ssi. *The Quarterly Journal of Economics*, 137(4):2263–2307. 3, 4

Diamond, P. and Sheshinski, E. (1995). Economic aspects of optimal disability benefits. *Journal of Public economics*, 57(1):1–23. 3

Dobkin, C. and Puller, S. L. (2007). The effects of government transfers on monthly cycles in drug abuse, hospitalization and mortality. *Journal of Public Economics*, 91(11-12):2137–2157. 39

Doyle Jr, J. J. (2007). Child protection and child outcomes: Measuring the effects of foster care. *American Economic Review*, 97(5):1583–1610. 13

Duflo, E., Dupas, P., and Kremer, M. (2024). The impact of secondary school subsidies on career trajectories in a dual labor market: Experimental evidence from ghana. CEPR Discussion Paper 19344, Centre for Economic Policy Research. 4

Dustmann, C. and Meghir, C. (2005). Wages, experience and seniority. *The Review of Economic Studies*, 72(1):77–108. 39

Dustmann, C., Mertz, M., and Okatenko, A. (2023). Neighbourhood gangs, crime spillovers and teenage motherhood. *The Economic Journal*, 133(653):1901–1936. 24

Dynarski, S. M. (2003). Does aid matter? measuring the effect of student aid on college attendance and completion. *American Economic Review*, 93(1):279–288. 3

Eurostat (2026). Young people – health. Statistics Explained. Data extracted June 2026, article updated July 7, 2026. 1

Evans, W. N. and Moore, T. J. (2011). The short-term mortality consequences of income receipt. *Journal of Public Economics*, 95(11-12):1410–1424. 2, 39, 43

Filippucci, F. (2025). What do neets need? the overall effect of active and passive labor market policies. *Working Paper*. 4, 39, 46

Finkelstein, A. and Notowidigdo, M. J. (2019). Take-up and targeting: Experimental evidence from snap. *The Quarterly Journal of Economics*, 134(3):1505–1556. 43

Fogelgren, M., Ornstein, P., Rödin, M., and Skogman Thoursie, P. (2023). Is supported employment effective for young adults with disability pension? evidence from a swedish randomized evaluation. *Journal of Human Resources*, 58(2):452–487. 4

Frandsen, B., Lefgren, L., and Leslie, E. (2023). Judging judge fixed effects. *American Economic Review*, 113(1):253–277. 18, 20

Frandsen, B. R., Lefgren, L. J., Leslie, E. C., and McIntyre, S. P. (2026). The incredible shrinking instruments: Using empirical bayes to increase efficiency in iv designs with many instruments. Working Paper 34634, National Bureau of Economic Research. 15, 89

French, E. and Song, J. (2014). The effect of disability insurance receipt on labor supply. *American economic Journal: economic policy*, 6(2):291–337. 3, 13

Gelber, A., Isen, A., and Kessler, J. B. (2016). The effects of youth employment: Evidence from new york city lotteries. *The Quarterly Journal of Economics*, 131(1):423–460. 42

Gelber, A., Moore, T., Pei, Z., and Strand, A. (2023). Disability insurance income saves lives. *Journal of Political Economy*, 131(11):3156–3185. 3, 4

Glewwe, P. and Kassouf, A. L. (2012). The impact of the bolsa escola/familia conditional cash transfer program on enrollment, dropout rates and grade promotion in brazil. *Journal of Development Economics*, 97(2):505–517. 26

Goffman, E. (1963). *Stigma: Notes on the Management of Spoiled Identity*. Prentice-Hall, Englewood Cliffs, NJ. 42

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2026). Leniency designs: An operator's manual. *Journal of Economic Perspectives*, 40(3):213–240. 32

Haaga, T., Böckerman, P., Kortelainen, M., and Tukiainen, J. (2024). Effects of nurse visit copayment on primary care use: Do low-income households pay the price? *Journal of Health Economics*, 94:102866. 30

Haller, A., Staubli, S., and Zweimüller, J. (2024). Designing disability insurance reforms: Tightening eligibility rules or reducing benefits? *Econometrica*, 92(1):79–110. 3

Heckman, J., Pinto, R., and Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American economic review*, 103(6):2052–2086. 40

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to

evaluate social programs, and to forecast their effects in new environments. *Handbook of econometrics*, 6:4875–5143. 20

Heeks, M., Reed, S., Tafsiri, M., and Prince, S. (2018). The economic and social costs of crime. Research Report 99, Home Office, London. 85

Helénsdotter, R. (2026). Surviving childhood: Effects of removing a child from home. *Review of Economic Studies*, 93(2):1001–1037. 13

Heller, S. B. (2014). Summer jobs reduce violence among disadvantaged youth. *Science*, 346(6214):1219–1223. 26, 42

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *The Quarterly journal of economics*, 135(3):1209–1318. 3, 43, 44, 45

Hensvik, L., Müller, D., and Skans, O. N. (2023). Connecting the young: High school graduates' matching to first jobs in booms and great recessions. *The Economic Journal*, 133(652):1466–1509. 4, 26, 42

Hinkkanen, V. (2014). Rikollisuuden kustannukset. In Lappi-Seppälä, T. and Kivivuori, J., editors, *Rikollisuustilanne 2013: Rikollisuus ja seuraamusjärjestelmä tilastojen valossa*, number 266 in Oikeuspoliittisen tutkimuslaitoksen tutkimuksia, pages 411–420. Oikeuspoliittinen tutkimuslaitos, Helsinki. 85

Hotz, V. J., Xu, L., Tienda, M., and Ahituv, A. (2002). Are there returns to the wages of young men from working while in school? *The Review of Economics and Statistics*, 84(2):221–236. 42

Hull, P. (2017). Examiner designs and first-stage f statistics: A caution. *Unpublished Working Paper*, 5. 16

Humlum, A., Munch, J. R., and Plato, P. (2025). Changing tracks: human capital investment after loss of ability. *American Economic Review*, 115(5):1520–1554. 4

Humlum, M. K. and Vejlin, R. (2013). The responses of youth to a cash transfer conditional on schooling: A quasi-experimental study. *Journal of Applied Econometrics*, 28(4):628–649. 4

Humphries, J. E., Ouss, A., Stavreva, K., Stevenson, M. T., and van Dijk, W. (2025). Conviction, incarceration, and recidivism: Understanding the revolving door. *The Quarterly Journal of Economics*, 140(4):2907–2962. 15, 18, 89

Huynh, J. and Chan, D. (2024). Do cash transfers narrow health disparities? evidence from veterans with disabilities. *Working Paper*. 3, 4

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475. 18

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York. 21

Imbens, G. W. and Xu, Y. (2025). Comparing experimental and nonexperimental methods: What lessons have we learned four decades after lalonde (1986)? *Journal of Economic Perspectives*, 39(4):173–201. 21

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876. 13

Kluve, J., Puerto, S., Robalino, D., Romero, J. M., Rother, F., Stöterau, J., Weidenkaff, F., and Witte, M. (2019). Do youth employment programs improve labor market outcomes? a quantitative review. *World Development*, 114:237–253. 4

Knutsson, D. and Tyrefors, B. (2024). Labor market effects of a youth summer employment program in sweden. Technical report, Working Paper. 26, 42

Kolesar, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Technical report. 32

Koskenvuo, K., Ryynänen, M., Kemppinen, H., Kokkonen, V., and Autti-Rämö, I. (2014). Vajaakuntoiselle nuorelle tarvitaan selkeämpi reitti ammattitutkintoon. *Sosiaalivakuutus*, (1):35–36. 7

Kostøl, A. R. and Mogstad, M. (2015). Earnings, disposable income, and consumption of allowed and rejected disability insurance applicants. *American Economic Review*, 105(5):137–141. 3, 13

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620. 21

Le Barbanchon, T., Ubfal, D., and Araya, F. (2020). The effects of working while in school: evidence from uruguayan lotteries. *Available at SSRN 3398385*. 4, 42

Levere, M. (2021). The labor market consequences of receiving disability benefits during childhood. *Journal of Human Resources*, 56(3):850–877. 1, 3

Lindbeck, A., Nyberg, S., and Weibull, J. W. (2003). Social norms and welfare state dynamics. *Journal of the European Economic Association*, 1(2-3):533–542. 39

Lindquist, M. J. and Santavirta, T. (2014). Does placing children in foster care increase their adult criminality? *Labour Economics*, 31:72–83. 23, 24

Maczulskij, T., Kanninen, O., Karhunen, H., and Tahvonen, O. (2025). Debt burden of job loss in a nordic welfare state. *Working Paper*. 10

Maestas, N., Mullen, K. J., and Strand, A. (2013). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American economic review*, 103(5):1797–1829. 1, 3, 13, 18, 20, 38, 46, 73

Markussen, S. and Røed, K. (2014). The impacts of vocational rehabilitation. *Labour Economics*, 31:1–13. 4

Mian, A., Sufi, A., and Khoshkhou, N. (2023). Partisan bias, economic expectations, and household spending. *Review of Economics and Statistics*, 105(3):493–510. 24

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, 66(4):281–302. 39

Mincer, J. A. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research, New York. 39

Mirrlees, J. A. (1971). An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208. 46

Moffitt, R. (1983). An economic model of welfare stigma. *The American economic review*, 73(5):1023–1035. 39, 42

Mogstad, M., Santos, A., and Torgovitsky, A. (2018). Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619. 18

Mogstad, M. and Torgovitsky, A. (2024). Instrumental variables with unobserved heterogeneity in treatment effects. In *Handbook of Labor Economics*, volume 5, pages 1–114. Elsevier. 20, 32, 73

Nordström Skans, O. (2011). Scarring effects of the first labor market experience. Technical Report 5565, Institute for the Study of Labor (IZA). 4

Norris, S., Pecenco, M., and Weaver, J. (2021). The effects of parental and sibling incarceration: Evidence from ohio. *American Economic Review*, 111(9):2926–2963. 20

Nunn, N. and Wantchekon, L. (2011). The slave trade and the origins of mistrust in africa. *American economic review*, 101(7):3221–3252. 24

Opetushallitus (2025). Koulutuksen kustannukset kasvoivat edellisvuosia vähemmän vuonna 2024. https://www.oph.fi/fi/uutiset/2025/koulutuksen-kustannukset-kasvoivat-edellisvuosia-vahemman-vuonna-2024. Accessed August 20, 2026. 86

Oreopoulos, P., Von Wachter, T., and Heisz, A. (2012). The short-and long-term career effects of graduating in a recession. *American Economic Journal: Applied Economics*, 4(1):1–29. 4

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204. 21, 24, 34

Parker, S. W. and Todd, P. E. (2017). Conditional cash transfers: The case of progresa/oportunidades. *Journal of Economic Literature*, 55(3):866–915. 46

Parker, S. W. and Vogl, T. (2023). Do conditional cash transfers improve economic outcomes in the next generation? evidence from mexico. *The Economic Journal*, 133(655):2775–2806. 4

Price, D. J. and Song, J. (2025). The long-term effects of cash assistance. *American Economic Journal: Applied Economics*, Forthcoming. 39

Raaum, O. and Røed, K. (2006). Do business cycle conditions at the time of labor market entry affect future employment prospects? *The Review of Economics and Statistics*, 88(2):193–210. 4

Reiterä, T. (2021). Myönteisen kuntoutuspäätöksen vaikutus nuoren työikäisen sosiaaliturvan tarpeeseen. *Master's thesis*. 4, 16

Ruhm, C. J. (1997). Is high school employment consumption or investment? *Journal of Labor Economics*, 15(4):735–776. 42

Salminen, J. (2019). Nuoren kuntoutusrahan luoma osallisuus toimintamahdollisuuksina. *Yhteiskuntapolitiikka*, 84. 43

Schultz, T. P. (2004). School subsidies for the poor: Evaluating the mexican progresa poverty program. *Journal of Development Economics*, 74(1):199–250. 26

Sigstad, H. (2024). Marginal treatment effects and monotonicity. *arXiv preprint arXiv:2404.03235*. 20, 73

Silver, D. and Zhang, J. (2022). Invisible wounds: health and well-being impacts of mental disorder disability compensation on veterans. Technical report, National Bureau of Economic Research. 3, 4

Smith, J. A. and Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353. 21

Statistics Finland (2025a). EDUC OPISK: Opiskelijat. https://taika.stat.fi/fi/aineistokuvaus.html#!?dataid=EDUC_19992023_jua_opisk_001.xml. 9

Statistics Finland (2025b). EDUC TYHR: Toisen asteen yhteishaku. https://taika.stat.fi/fi/aineistokuvaus.html#!?dataid=EDUC_20142023_jua_tyhr24_001.xml. 9

Statistics Finland (2025c). Folk laps. https://taika.stat.fi/fi/aineistokuvaus.html#!?dataid=FOLK_19702023_jua_laps24_001.xml. 9

Statistics Finland (2025d). Folk perus. https://taika.stat.fi/fi/aineistokuvaus.html#!?dataid=FOLK_19872023_jua_perus24_001.xml. 9

Valtioneuvosto (2024). Hallituksen esitys eduskunnalle laiksi rikosasioiden ja eräiden riita-asioiden sovittelusta annetun lain muuttamisesta. Hallituksen esitys. Finlex. 85

Valtioneuvosto (2026). Hallituksen esitys eduskunnalle laiksi rikoslain 25 luvun muuttamisesta ja eräiksi muiksi laeiksi. Hallituksen esitys. Finlex. 85

Appendix to

Cash Transfers, Health, and the Transition to Adulthood

Ossi Tahvonen

Contents

A.1 Institutional details 57
A.2 Additional Results 58
A.2.1 Education, health, and high-risk-group outcomes 58
A.2.2 Transfer receipt and public costs 63
A.2.3 Covariate-adjusted estimates and treatment-effect heterogeneity 65
A.2.4 Alternative treatment definitions, reduced forms, and mechanisms 68
A.2.5 Additional descriptive evidence 70
A.3 Additional validity and robustness checks 73
A.3.1 Monotonicity 73
A.3.2 Robustness 73
A.4 Marginal Treatment Effects 78
A.5 Derivations of the human capital model 79
A.6 Methodological Details 81
A.6.1 Estimation of effects on parents using differences-in-differences/event study 81
A.6.2 Selection-on-observables approach 81
A.6.3 Overlap and placebo diagnostics for the high-DI-risk subgroup 82
A.6.4 Additional DDML-IV comparisons 84
A.6.5 Cost-effectiveness and MVPF calculations 85
A.6.6 Empirical Bayes shrinkage and simulation evidence 87

A.1 Institutional details

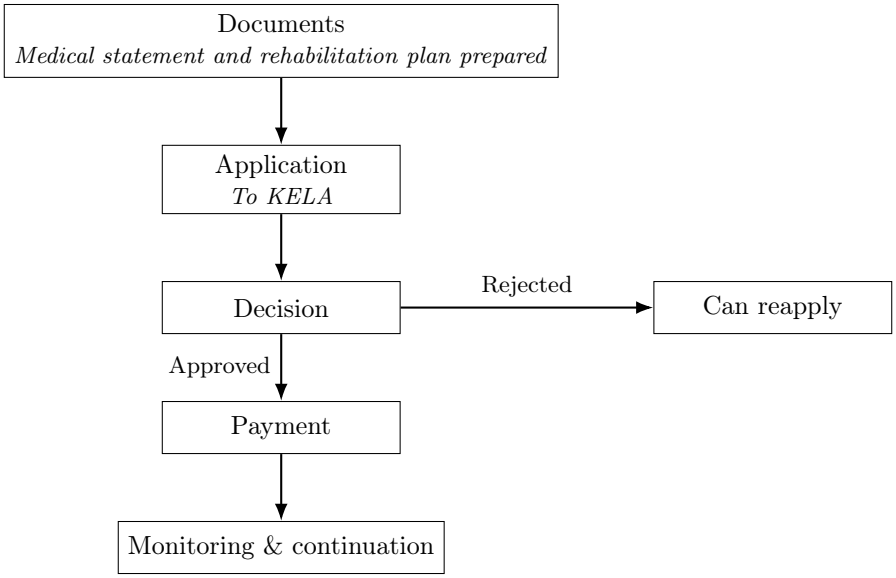
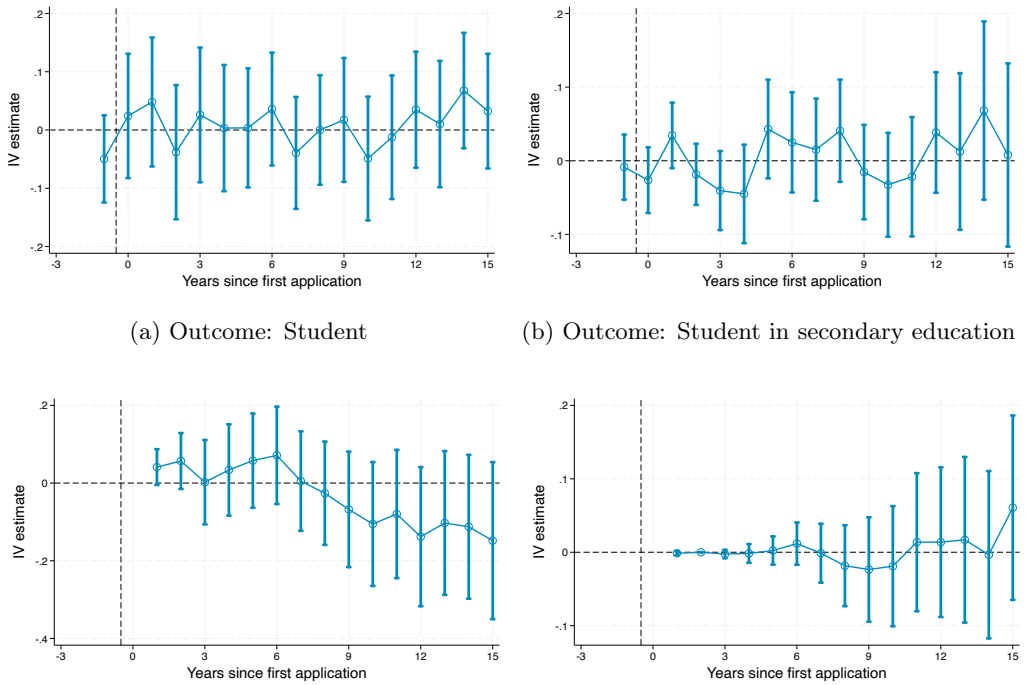


Figure A.1: Application process for the Youth Rehabilitation Allowance.

A.2 Additional Results

A.2.1 Education, health, and high-risk-group outcomes



(a) Outcome: Student

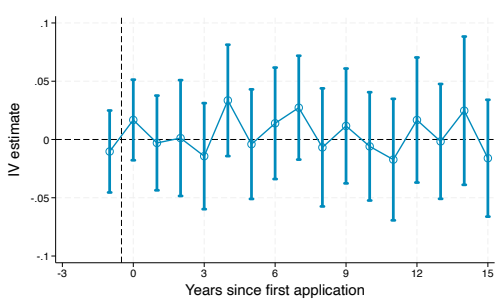
(b) Outcome: Student in secondary education

(c) Outcome: Secondary education degree

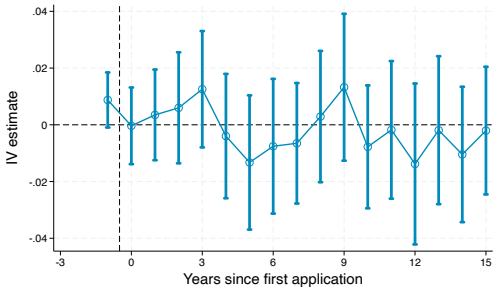
(d) Outcome: Higher education degree

Figure A.2: Long-run (EB) IV estimates on educational outcomes.

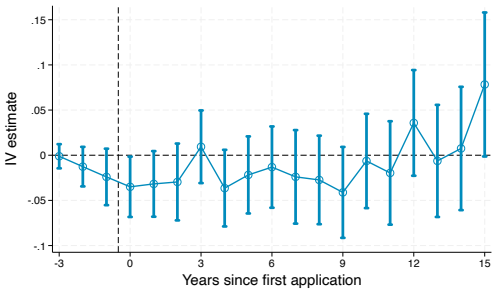
Notes: Each estimate represents an IV-estimate from a separate regression using time-specific samples. Outcomes apart from health are not necessarily informative for individuals younger than 16 years old, which moves pre-treatment estimates toward 0. Estimates are not scaled, so in the case of binary outcomes an estimate of 0.1 means 10p.p. increase.



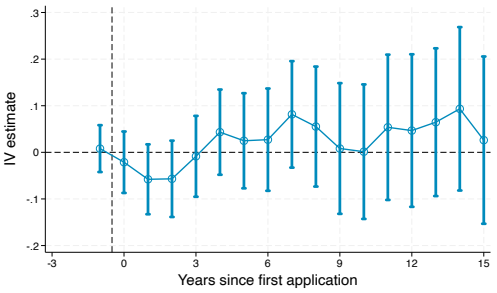
(a) Outcome: Any criminal conviction



(b) Outcome: Any drug related criminal conviction



(c) Any addiction diagnosis



(d) Outcome: Financial trouble (garnishment)

Figure A.3: Long-run (EB) IV estimates on crime, drug-related crimes, addiction diagnoses, and financial trouble.

Notes: Each estimate represents an IV-estimate from a separate regression using time-specific samples. Outcomes apart from health are not necessarily informative for individuals younger than 16 years old, which moves pre-treatment estimates toward 0. Estimates are not scaled, so in the case of binary outcomes an estimate of 0.1 means 10p.p. increase.

	(1) Hospital days	(2) Overnight MH visits	(3) Any MH visit	(4) Any antidepressants
Panel A. Main IV estimates				
	0.024	**0.005**	**-0.070**	**-0.029**
	(1.105)	**(0.064)**	**(0.044)**	**(0.046)**
AR p-value	0.983	0.941	0.112	0.521
First-stage F	343	343	343	343
Panel B. IV with additional controls				
	-0.087	-0.003	-0.077*	-0.030
	(1.104)	(0.064)	(0.043)	(0.045)
AR p-value	0.937	0.965	0.068	0.512
First-stage F	346	346	346	346
Panel C. OLS benchmark				
	0.092	0.011**	0.004	-0.010***
	(0.065)	(0.005)	(0.003)	(0.003)
Untreated complier mean	1.84	0.127	0.346	0.261
Sample mean	1.56	0.106	0.341	0.239
Observations	117,649	117,649	117,649	117,649
Individuals	26,617	26,617	26,617	26,617

Table A.1: Main IV results on health and healthcare use in years 6–12 after the first application.

Notes: AR denotes the Anderson–Rubin test. Outcomes are measured years 6-12 after the first application, which roughly corresponds to ages 21-28. Health outcomes are available until 2023. Both OLS and IV use the same, necessary controls. IV with additional controls uses controls for the disability category, indicators for any depression/anxiety/F9/other healthcare visits, urban area, parental employment, parent ages, and 9th grade GPA. Sample size refers to individual × year observations. Standard errors clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	(1) Cumulative schooling (years)	(2) Secondary degree	(3) Cumulative DI receipt (years)	(4) Cumulative wages (euros)
Panel A. Main IV estimates				
	-0.089	**-0.027**	**-2.031*****	**-664**
	(0.618)	**(0.174)**	**(0.574)**	**(2,546)**
AR p-value	0.885	0.876	0.002	0.794
First-stage F	39.1	39.1	39.1	39.1
Panel B. IV with additional controls				
	-0.054	-0.002	-2.019***	-475
	(0.614)	(0.172)	(0.572)	(2,526)
AR p-value	0.929	0.993	0.002	0.851
First-stage F	39.5	39.5	39.5	39.5
Panel C. OLS benchmark				
	0.673***	0.059***	-2.042***	-185
	(0.049)	(0.014)	(0.048)	(226)
Untreated complier mean	2.7	0.204	2.53	2,789
Sample mean	3.33	0.25	1.24	2,054
Individuals	5,724	5,720	5,724	5,724

Table A.2: Main IV results on cumulative education and labor-market outcomes by age 21 for the high DI risk sample.

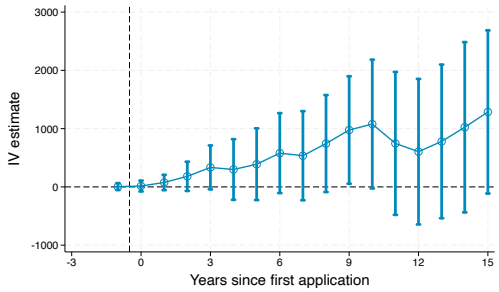
Notes: AR denotes the Anderson–Rubin test. Euro amounts are rounded to the nearest euro. Both OLS and IV use the same, necessary controls. IV with additional controls uses controls for the disability category, indicators for any depression/anxiety/F9/other healthcare visits, urban area, parental employment, parent ages, and 9th grade GPA. Standard errors clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

	(1) NEET	(2) Employed	(3) Wage income (euros)	(4) Student
Panel A. Main IV estimates				
	0.051	**-0.094**	**-2,839**	**0.040**
	(0.182)	**(0.160)**	**(3,416)**	**(0.094)**
AR p-value	0.781	0.556	0.402	0.669
First-stage F	24.5	24.5	24.5	24.5
Panel B. IV with additional controls				
	0.018	-0.066	-2,395	0.045
	(0.178)	(0.156)	(3,335)	(0.093)
AR p-value	0.920	0.671	0.470	0.624
First-stage F	25.1	25.1	25.1	25.1
Panel C. OLS benchmark				
	-0.034***	0.031***	93	0.003
	(0.007)	(0.006)	(123)	(0.004)
Untreated complier mean	0.763	0.211	5,914	0.0248
Sample mean	0.77	0.157	2,826	0.0725
Observations	22,997	22,997	22,997	22,997
Individuals	4,892	4,892	4,892	4,892

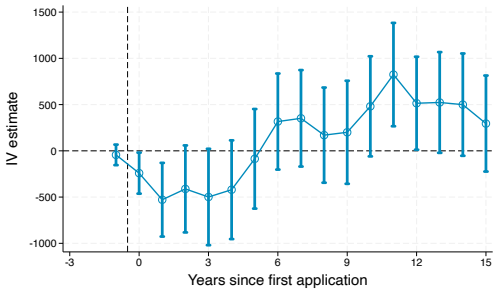
Table A.3: Main IV results on labor-market outcomes in years 6–11 after the first application for the high DI risk sample.

Notes: AR denotes the Anderson–Rubin test. Euro amounts are rounded to the nearest euro. Outcomes are measured years 6-11 after the first application, which roughly corresponds to ages 21-28. Both OLS and IV use the same, necessary controls. IV with additional controls uses controls for the disability category, indicators for any depression/anxiety/F9/other healthcare visits, urban area, parental employment, parent ages, and 9th grade GPA. Sample size refers to individual × year observations. Standard errors clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

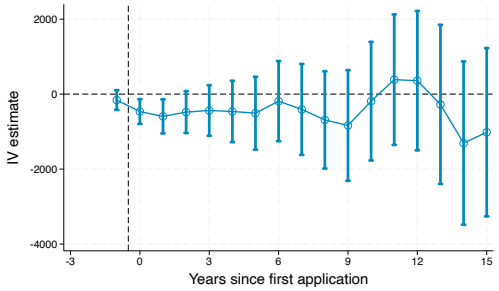
A.2.2 Transfer receipt and public costs



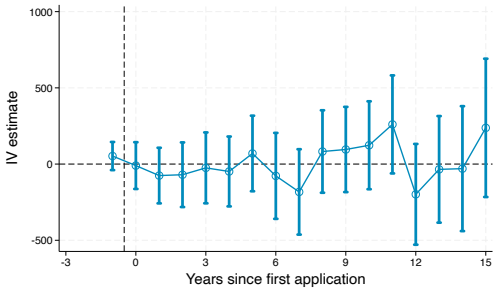
(a) Outcome: Unemployment benefits (euros)



(b) Outcome: Social Assistance (euros)



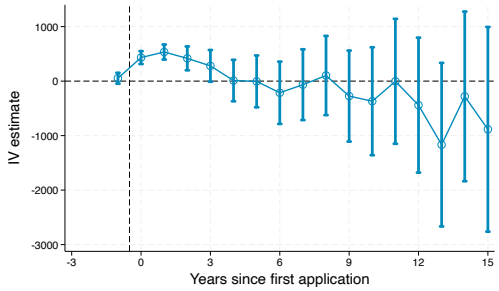
(c) Outcome: Pensions paid (euros) (disability insurance)



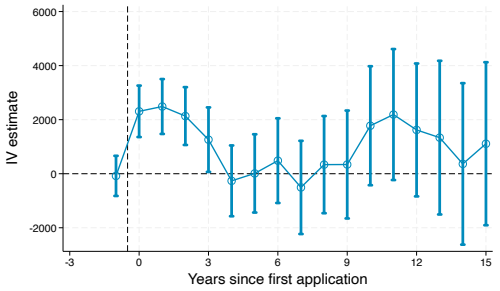
(d) Outcome: Illness benefits (euros)

Figure A.4: Long-run (EB) IV estimates on benefits paid to individuals.

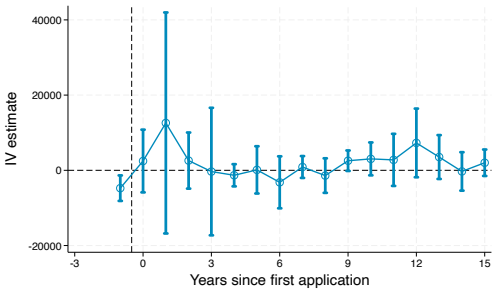
Notes: Each estimate represents an IV-estimate from a separate regression using time-specific samples. Outcomes apart from health are not necessarily informative for individuals younger than 16 years old, which moves pre-treatment estimates toward 0.



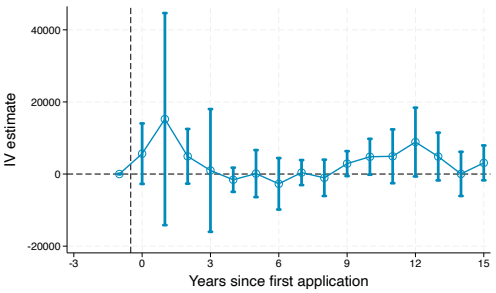
(a) Outcome: Paid taxes (euros)



(b) Outcome: Total transfers (euros)



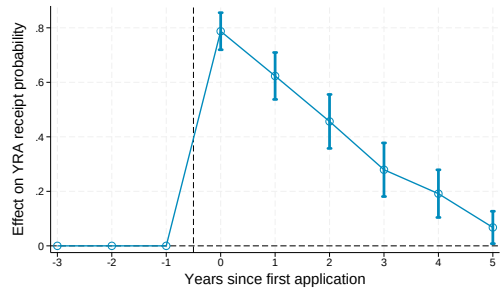
(c) Outcome: Hospital and specialized care costs (euros)



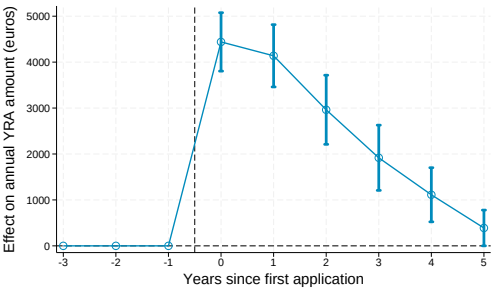
(d) Outcome: All public health expenses (euros)

Figure A.5: Long-run (EB) IV estimates on cumulative costs to the public sector.

Notes: Each estimate represents an IV-estimate from a separate regression using time-specific samples.



(a) Outcome: Currently in YRA

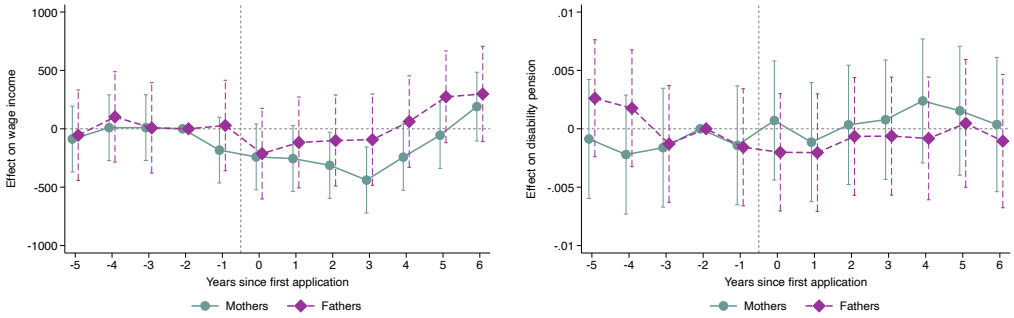


(b) Outcome: Amount of YRA received

Figure A.6: IV estimates on receipt of YRA and the amount received each year.

Notes: Each estimate represents an IV-estimate from a separate regression using time-specific samples.

A.2.3 Covariate-adjusted estimates and treatment-effect heterogeneity



(a) Outcome: Wage income

(b) Outcome: Disability Insurance

Figure A.7: Effects on parents labor market outcomes and social insurance use.

Notes: Estimates using the DiD specification in Section A.6.1. Wage income is the income from wages in 2025 euros, Disability Insurance is an indicator whether the individual received Disability Insurance during that year. All applicants included.

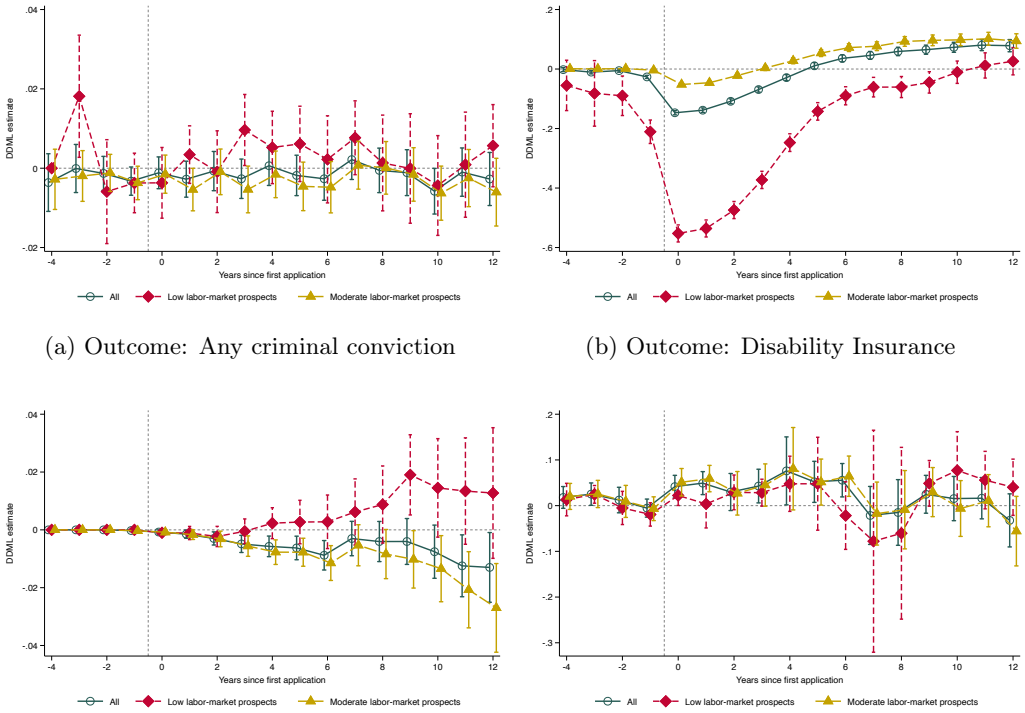
	(1) Applied for YRA	(2) N of applications	(3) Received YRA
Approved	0.052	0.136**	0.057**
	(0.035)	(0.055)	(0.023)
Sample mean	.06	.083	.057
Controls	IV controls	IV controls	IV controls
Ages in sample	15-17	15-17	15-22
N	28787	28787	87768
Individuals	14115	14115	29184

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Table A.4: IV estimates on the YRA applications and received YRA for the siblings of the original applicants.

Notes: A positive coefficient indicates that the original applicant being approved to YRA increases the applications/recipiency of their siblings. Standard errors clustered by individual (applicant). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.



(a) Outcome: Any criminal conviction

(b) Outcome: Disability Insurance

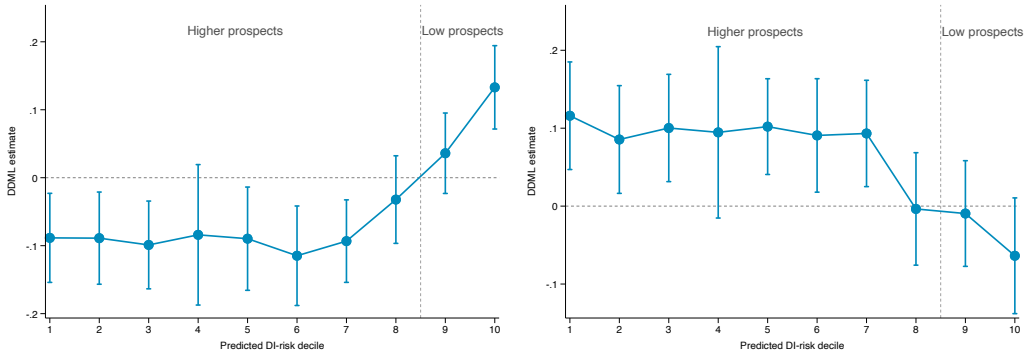
(c) Outcome: Married

(d) Outcome: Overnight MH visit

Figure A.8: Additional results using the covariate adjustment approach.

Notes: Each estimate represents a DDML estimate from a separate estimation using time-specific samples. Marriage is not possible for individuals younger than 18 years old, which is why the pre-treatment outcomes are not informative. Estimates are not scaled, so in the case of binary outcomes, an estimate of 0.1 means 10p.p. increase. 'All' refers to the entire sample with overlap, 'Low Labor Market prospects' to the applicants with the lowest labor market prospects and 'Moderate Labor Market prospects' to the rest.

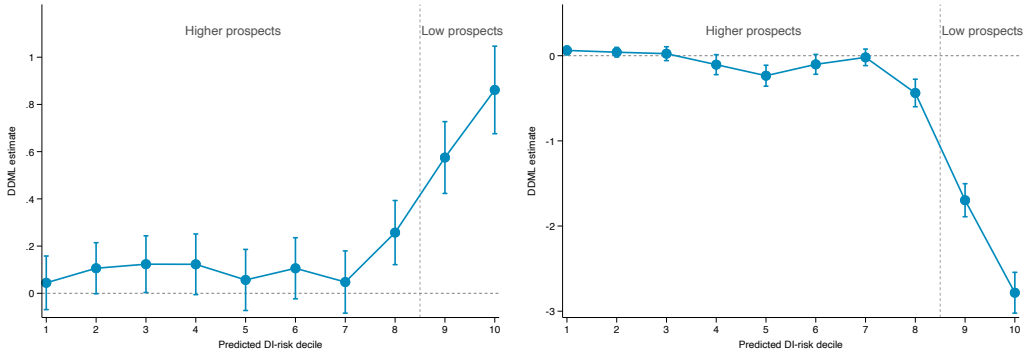
Adult outcomes, pooled over years 6–12 after first application



(a) Outcome: Employed

(b) Outcome: NEET

Cumulative outcomes by age 21



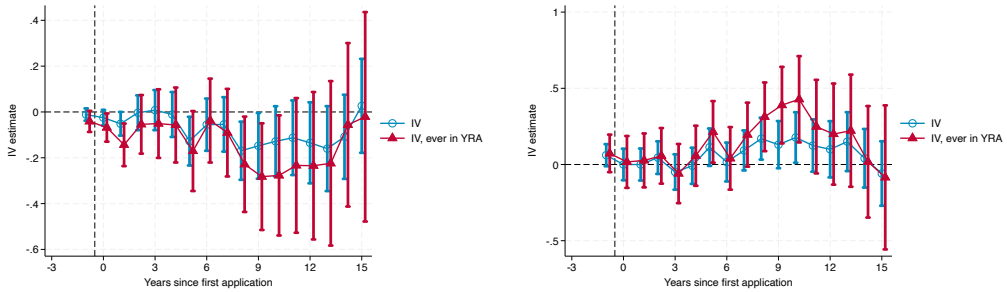
(c) Outcome: Years in education

(d) Outcome: Years on DI

Figure A.9: DDML estimates by predicted Disability Insurance (DI) risk.

Notes: The index predicts DI receipt two years after application using pre-treatment covariates and is estimated out of fold among rejected applicants. Applicants are grouped into deciles ordered from higher to lower labor-market prospects. The dashed vertical line separates the top fifth of predicted DI risk (deciles 9–10), which defines the low-labor-market-prospect group in the main analysis, from the remaining applicants. Each point is a separate DDML estimate for the corresponding decile using the specification described in Section 4.3. Vertical bars show 95% confidence intervals. Estimates are not scaled, so an estimate of 0.1 for a binary outcome denotes a 10 percentage-point effect. Standard errors are clustered at the individual level.

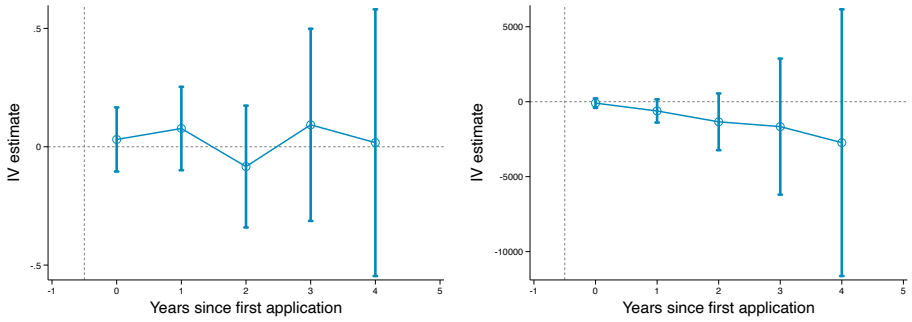
A.2.4 Alternative treatment definitions, reduced forms, and mechanisms



(a) Outcome: Employed (b) Outcome: NEET

Figure A.10: Results using 'first application accepted' as the endogenous variable and a comparison to using 'Ever Receive any YRA benefits' as the endogenous variable.

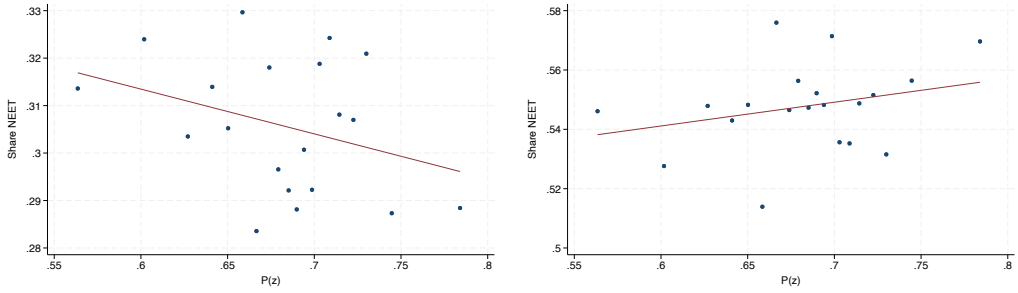
Notes: 95 % confidence intervals included.



(a) Outcome: Student (b) Outcome: Wage earnings

Figure A.11: IV results using 'Currently in the YRA program' as the endogenous variable.

Notes: 95 % confidence intervals included.



(a) Outcome: Employed (b) Outcome: NEET

Figure A.12: Reduced-form graphs.

Notes: The x-axis is the first-stage predicted approval probability, $P(z)$, and the y-axis is the outcome measured 6–11 years after application.

	(1) NEET	(2) Employed	(3) Wage inc.
Reduced form	0.0955**	-0.0911**	-1430.3
	(0.0444)	(0.0410)	(1092.3)
Observations	102592	102592	102592

Standard errors in parentheses

* p<0.1, ** p<0.05, *** p<0.01

Table A.5: Reduced form version of the main results (as in Table 6).

	(1) IV	(2) IV, mediation	(3) DDML	(4) DDML, mediation	(5) DDML (compliers)	(6) DDML (compliers), mediation
First applic. accepted	-0.112**	-0.0936**	-0.0461***	-0.00618**	-0.0930***	-0.0224***
	(0.0500)	(0.0424)	(0.00303)	(0.00274)	(0.00676)	(0.00615)
Outcome mean	0.36	0.36	0.29	0.29	0.36	0.36
Controls	IV controls	IV controls + cum. earnings	All	All + cum. earnings	All	All + cum. earnings
Observations	102592	102592	102595	102595	102595	102595

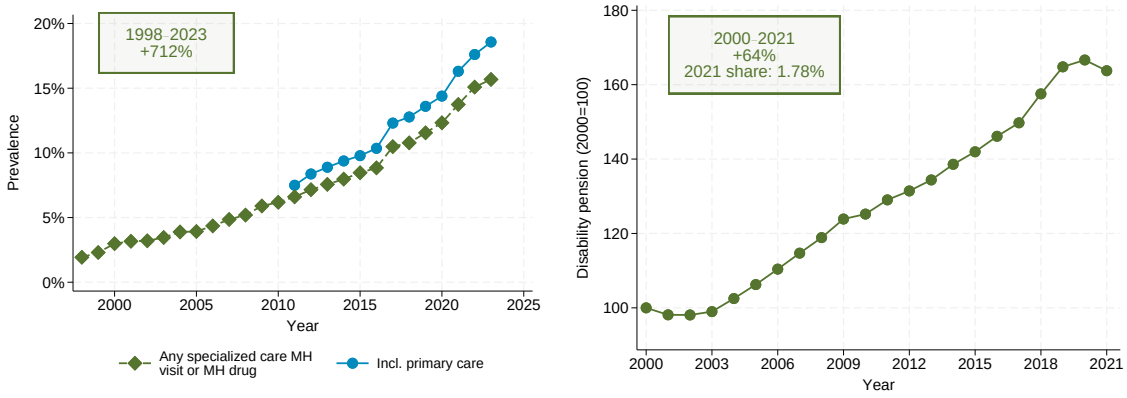
Standard errors in parentheses

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

Table A.6: Mediation analysis.

Notes: Different specifications with and without controlling for accumulated wage earnings at age 20. The outcome is employment 6-12 years from the application. Since the mediator is not randomized and there is no instrument for it available, the results should be taken as suggestive. Standard errors clustered at the individual level.

A.2.5 Additional descriptive evidence



(a) Share with some mental health treatment contact, with and without primary care.

(b) Disability-pension recipients, ages 16–29.

Figure A.13: Yearly disability pension receipt among individuals aged 16–29 (panel b) and MH treatment contact among individuals aged 13–17 (panel a) in Finland.

Notes: Each series is indexed to 100 in 2000. Sources: Finnish Centre for Pensions (ETK), Statistics Finland, author's own calculations using HILMO, Avohilmo, KELA drug reimbursement data.

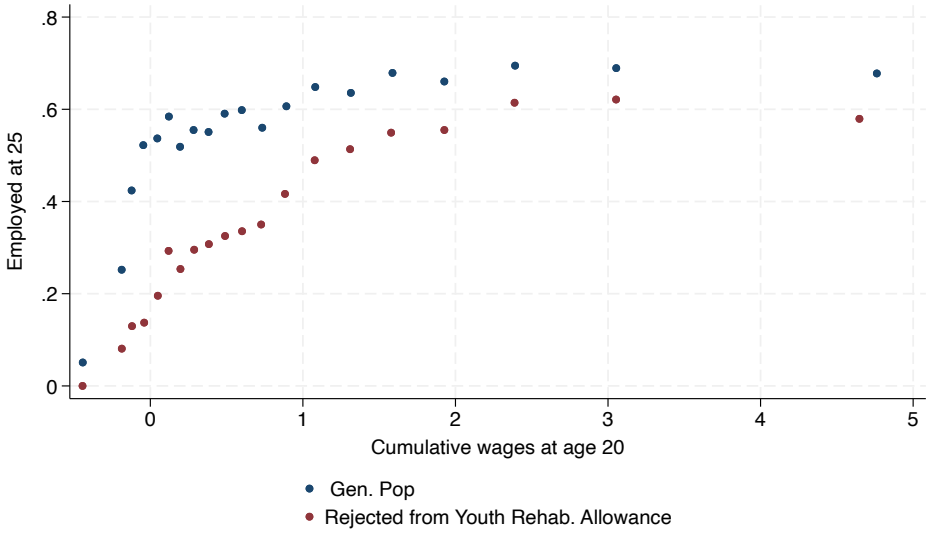


Figure A.14: Descriptive figure of the association between accumulated work earnings (x-axis) and the probability of employment (y-axis) between years 6–12 after the application to YRA.

Notes: It displays 20 even-sized bins. 'Gen. Pop.' refers to a group with a similar cohort profile as the applicants for comparison.

	(1) Gen. pop.	(2) Approved	(3) Rejected
Student	0.976	0.997	0.998
Employed	0.024	0.003	0.002
Distinct admissions	1.02	3.85	3.85
Healthcare costs	559	2,424	2,164
Anxiety diagnosis	0.053	0.019	0.020
Depression diagnosis	0.052	0.015	0.017
Single parent HH	0.211	0.246	0.253
Urban	0.807	0.747	0.762
Mother age	44.04	44.51	44.80
Father age	46.45	47.60	47.69
Mother DI	0.027	0.102	0.099
Father DI	0.040	0.091	0.088
Mother married	0.697	0.577	0.598
Father married	0.691	0.563	0.577
Mother employed	0.814	0.652	0.691
Father employed	0.831	0.707	0.728
Mother taxable income	4,467	26	167
Father taxable income	6,174	40	290
Accepted (1st appl.)	0	1	0
Ever in YRA	0	0.996	0.242
Total YRA benefits	0	25,857	5,846
Observations	1,476,350	4,471	1,275

Table A.7: Summary statistics for the applicants, by approval status, for the low labor market prospect group.

Notes: All variables are measured in the year before the first application. Euro amounts are in 2023 prices. Gen. pop. is the sample of all other individuals aged 15.

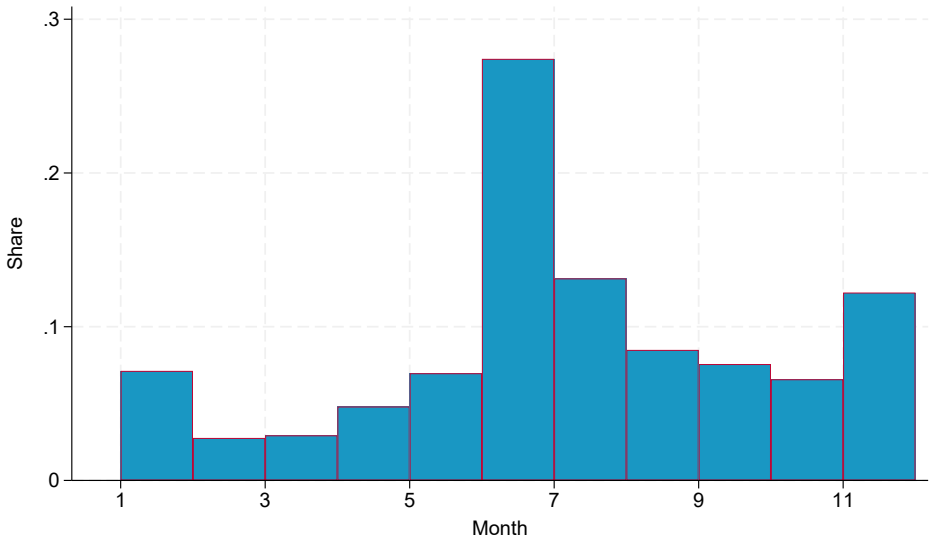


Figure A.15: Monthly distribution of working spells longer than 6 days among individuals receiving the Youth Rehabilitation Allowance during that year.

Notes: X-axis is the month, y-axis denotes the share of spells. The month is the starting month of the spell, since most spells are short, this corresponds to the end date as well. Data for years 2013-2020. This data is not available for all applicants.

Code	Diagnosis group	Approval rate	Share in DI after 5y	N
C7	Eye, brain/CNS, endocrine and unspecified cancers	79.6%	20.4%	54
C9	Leukemia, myeloma and other blood cancers	–	13.3%	45
D3	Benign neoplasms of urinary, nervous and endocrine organs	75.0%	25.0%	28
D6	Aplastic and other anemias	45.8%	–	24
E1	Diabetes mellitus and related glucose disorders	26.0%	4.0%	150
E2	Other endocrine-gland disorders	45.0%	–	20
E7	Inherited/metabolic disorders of amino acids, carbohydrates and lipids	65.6%	52.5%	61
E8	Other metabolic disorders	74.3%	–	35
F0	Organic mental disorders, including dementia	40.5%	13.5%	37
F1	Mental and behavioral disorders due to substance use	52.9%	–	34
F2	Schizophrenia, schizotypal and delusional disorders	75.3%	40.1%	887
F3	Mood (affective) disorders	61.4%	11.0%	4,234
F4	Anxiety, stress-related and somatoform disorders	60.9%	9.2%	2,739
F5	Eating, sleep and other physiologically linked behavioral syndromes	54.8%	11.1%	389
F6	Personality and adult behavioral disorders	61.0%	19.5%	77
F7	Intellectual disabilities	80.1%	55.8%	4,322
F8	Disorders of psychological development	72.0%	17.7%	7,823
F9	Childhood/adolescent-onset behavioral and emotional disorders	66.7%	7.2%	4,041
G1	Systemic atrophies affecting the central nervous system	67.4%	30.4%	46
G3	Alzheimer disease and other degenerative nervous-system disorders	75.0%	28.6%	28
G4	Epilepsy, migraine, TIA, sleep and other episodic neurologic disorders	59.0%	18.1%	537
G6	Polyneuropathies and other peripheral nervous-system disorders	63.0%	–	54
G7	Neuromuscular-junction and muscle diseases	77.9%	36.2%	149
G8	Cerebral palsy and other paralytic syndromes	77.9%	32.9%	721
G9	Autonomic and other nervous-system disorders	65.7%	25.7%	70
H2	Iris, ciliary-body and lens disorders	–	–	20
H3	Choroidal and retinal disorders	91.4%	21.6%	116
H4	Glaucoma, vitreous and optic-nerve disorders	86.7%	11.7%	60
H5	Eye-movement/refraction disorders and visual impairment	78.7%	13.8%	94
H9	Hearing loss and other ear disorders	79.6%	5.2%	574
I4	Cardiomyopathy, conduction disorders and cardiac arrhythmias	42.4%	–	33
I6	Cerebrovascular disease (stroke and related disorders)	63.9%	–	36
J4	Chronic lower respiratory diseases (including COPD and asthma)	21.5%	–	93
K5	Inflammatory bowel disease and other intestinal disorders	38.1%	–	84
L2	Dermatitis and eczema	12.9%	–	31
M0	Infectious arthropathies	47.5%	2.5%	510
M3	Systemic connective-tissue disorders	53.2%	–	47
M4	Spinal deformities and spondylopathies	37.5%	–	40
M5	Intervertebral-disc and other back disorders	19.4%	–	31
M8	Osteoporosis and other bone-density/structure disorders	38.1%	–	21
P1	Birth trauma	31.2%	–	32
Q0	Congenital malformations of the nervous system	79.8%	29.2%	168
Q1	Congenital malformations of eye, ear, face and neck	86.5%	14.6%	96
Q2	Congenital malformations of the circulatory system	56.3%	19.4%	103
Q6	Congenital malformations of urinary tract and selected musculoskeletal sites	64.8%	18.5%	54
Q7	Other congenital musculoskeletal malformations	67.8%	15.0%	180
Q8	Other congenital malformations	77.5%	39.0%	377
Q9	Chromosomal abnormalities	65.4%	73.8%	503
R4	Cognitive, consciousness, perception and behavioral symptoms/signs	40.3%	11.3%	62
R6	Edema, developmental and other general symptoms/signs	79.2%	29.2%	24
S0	Injuries of the head	55.8%	–	43
T9	Sequelae of injuries, poisoning and other external causes	42.9%	16.3%	49
Z9	Health status, follow-up and other factors influencing care	39.3%	–	28
–	Other (diagnosis groups with $N < 20$)	47.9%	2.4%	451

Table A.8: Diagnoses on the applications.

Notes: *Share approved* refers to the share of applications initially approved. *Share in DI after 5y* refers to the share of both rejected and accepted applicants who receive Disability Insurance/Pension at age 25. Due to data privacy, any shares that would correspond to less than 6 individuals have been changed to "–" in accordance with the Statistics Finland and Findata disclosure rules.

A.3 Additional validity and robustness checks

A.3.1 Monotonicity

This appendix describes the monotonicity tests summarized in Section 4.2.2 and reports their results.

The traditional test of monotonicity is to estimate the first stage for different subsamples. Monotonicity implies that the positive relationship between the instrument and the acceptance probability should hold in each sample (Maestas et al., 2013). Table A.9 reports the first-stage regression across subsamples.

Subsample	First-stage coefficient	Standard error	N
Male	0.785	0.058	16185
Female	0.819	0.065	14354
Urban	0.820	0.049	24353
Rural	0.689	0.094	6202
Mood disorder	0.887	0.121	4214
Anxiety disorder	0.964	0.153	2714
Physical condition	0.722	0.098	5964
Attention-deficit disorder	0.757	0.157	2375
Mother: higher education	0.758	0.080	9579
Mother: no higher education	0.826	0.052	20077
Decision year after 2007	0.791	0.053	23827
Finnish-speaking	0.820	0.044	29217

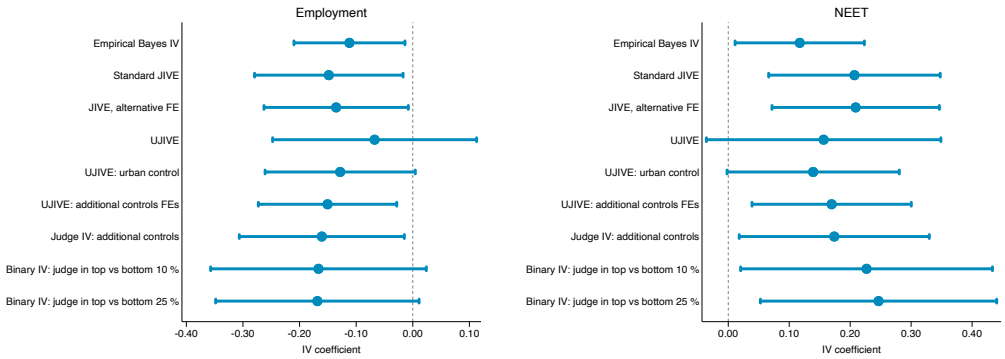
Table A.9: First-stage estimates by subgroup.

Notes: This is an indirect test of the monotonicity assumption: the first stage should be positive in all covariate groups.

Bhuller et al. (2020) propose a second test. An implication of the assumption is that a leniency measure estimated in one sample should be informative of the first stage in another sample. If the converse held, examiners who are lenient toward one group would be strict toward another, which would violate the assumption. I implement this test for groups based on the illness categories and gender in Table A.10. The first-stage coefficient is positive and significant in all splits, which supports the assumption.

As Mogstad and Torgovitsky (2024) point out, different constructions of the instrument might make the assumption of monotonicity more plausible. For instance, it is more likely that an examiner who is among the most lenient will be more lenient than an examiner who is among the strictest (Sigstad, 2024). I report results using this binary instrument in Fig. A.16. While the compliers using this instrument are plausibly different than with the continuous measure, the results are very similar to the main results.

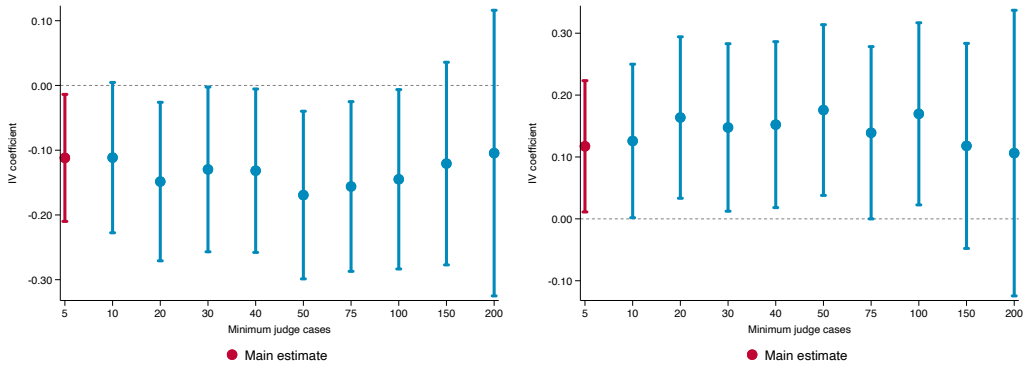
A.3.2 Robustness



(a) Employment (b) NEET

Figure A.16: Robustness to different IV estimation choices for the main outcomes.

Notes: 95 % confidence intervals included.



(a) Employment (b) NEET

Figure A.17: Robustness to different restrictions on the number of cases each examiner has to have handled.

Notes: The default with the Empirical Bayes shrunk IVs indicated in red. Outcomes measured at 6–11 years from the first application. Standard errors clustered at the individual level. 95 % confidence intervals included.

	First application approved	N	F
Female [Male]	0.412***	14301	83.05
	(0.0453)		
Male [Female]	0.365***	16080	88.72
	(0.0388)		
Others [Int./chrom. dis.]	0.117***	26549	48.93
	(0.0168)		
Others [Mood/anx.]	0.181***	22722	57.81
	(0.0238)		
Others [Dev. dis.]	0.296***	20562	71.52
	(0.0351)		
Others [Physical]	0.229***	18165	40.24
	(0.0361)		
Low acc. rate [High]	0.456***	9141	60.40
	(0.0587)		
High acc. rate [Low]	0.280***	21188	86.99
	(0.0300)		
Standard errors in parentheses			
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$			

Table A.10: Monotonicity test using the split-sample method (Bhuller et al., 2020).

Notes: Each row names the estimation group first and the group used to construct the instrument in brackets. For example, "Female [Male]" reports the first stage among female applicants using examiner leniency constructed from male applicants. The sample size refers to the estimation group.

	(1)	(2)	(3)
	First application approv.	First application approv.	First application approv.
EB IV	0.844***	0.803***	0.794***
	(0.0434)	(0.0433)	(0.0425)
N	30623	30623	30623
F-Statistic	378.8	344.4	348.7
Controls	No	Resid. FE	Res. + Additional
Standard errors in parentheses			
* p<0.1, ** p<0.05, *** p<0.01			

Table A.11: First stage estimates with different specifications.

Notes: *Resid. FE* refers to the necessary controls: the region-year-language FEs used in the construction of the IV. Additional controls include parent characteristics, diagnosis code group indicators, and number of healthcare visits before the application. The small difference between columns 2 and 3 is consistent with the instrument being independent of the background variables.

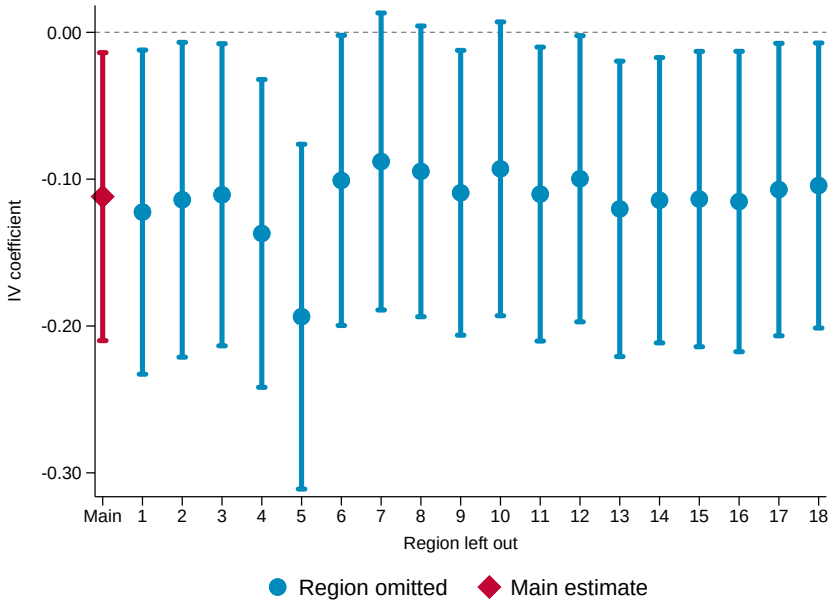


Figure A.18: Robustness of the employment estimate to excluding one region at a time.

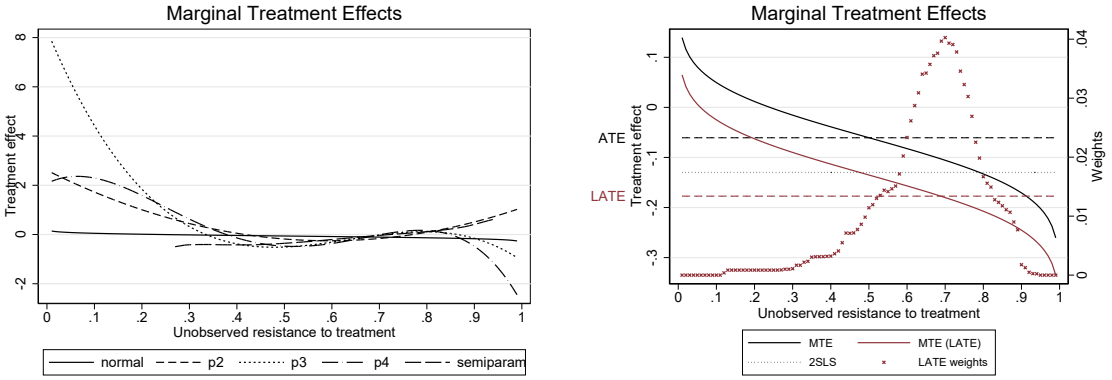
Notes: Blue circles show the leave-one-region-out EB IV estimates, while the red diamond shows the main full-sample estimate. Vertical bars indicate 95% confidence intervals.

	(1) NEET	(2) Employed	(3) Wage income (euros)	(4) Student
Panel A. Main IV estimates				
	0.160**	**-0.133****	**-1,853**	**-0.029**
	(0.066)	**(0.061)**	**(1,617)**	**(0.039)**
AR p-value	0.017	0.032	0.255	0.461
First-stage F	206	206	206	206
Panel B. IV with additional controls				
	0.148**	-0.124**	-1,736	-0.025
	(0.063)	(0.060)	(1,583)	(0.038)
AR p-value	0.019	0.038	0.274	0.520
First-stage F	208	208	208	208
Panel C. OLS benchmark				
	0.121***	-0.105***	-3,218***	-0.016***
	(0.004)	(0.003)	(86)	(0.003)
Untreated complier mean	0.443	0.407	9,309	0.15
Sample mean	0.585	0.283	6,416	0.131
Observations	77,901	77,901	77,901	77,901
Individuals	13,323	13,323	13,323	13,323

Table A.12: Main IV results on labor-market outcomes in years 6–11 after the first application for the balanced sample.

Notes: AR denotes the Anderson–Rubin test. Euro amounts are rounded to the nearest euro. The sample is limited to individuals observed each year from years 6 through 11 after the first application. IV with additional controls uses controls for the disability category, indicators for any depression/anxiety/F9/other healthcare visits, urban area, parental employment, parent ages, and 9th grade GPA. Sample size refers to individual × year observations. Standard errors clustered at the individual level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Marginal Treatment Effects



(a) MTE curves, using different parametric forms.

(b) MTE using normal distribution for the unobserved error term.

Figure A.19: Marginal treatment effects on employment.

Notes: In panel (a), 'pX' denotes polynomial degree X specification, normal denotes using normal distribution for the error terms. Outcome: employed, averaged over year 6-12 after application. LIV approach. Using the stata package `mtefe` (Andresen, 2018). Confidence intervals are not displayed due to computational as well as informational reasons: for most specifications, they span most of the possible interval for the binary outcome.

A.5 Derivations of the human capital model

Starting from the human capital equation: $H_i^A(d) = h(G_i) + \theta_S S_i(d) + g(X_i(d))$

we define that the change in human capital caused by the treatment is:

$\Delta H_i \equiv H_i^A(1) - H_i^A(0) = \theta_S \Delta S_i + \Delta g_i$.

Enter the definition of employment:

$E_i^A(d) = \mathbf{1}\{\log w_i^*(d) = \alpha + \lambda H_i^A(d) + u_i \geq R_i\}$. The treatment can cause changes in employment (and the parameter of interest is these changes in some population)

$\Delta E_i^A = \mathbf{1}\{\log w_i^*(1) \geq R_i\} - \mathbf{1}\{\log w_i^*(0) \geq R_i\}$.

Let $Z_i \equiv \log w_i^*(0) - R_i$. Conditional on a given treatment-induced change in human capital, $\Delta H_i = h$, the change in employment is

$$\Delta E_i = \mathbf{1}\{Z_i + \lambda h \geq 0\} - \mathbf{1}\{Z_i \geq 0\}.$$

Taking expectations conditional on $\Delta H_i = h$ gives

$$\begin{aligned}
\mathbb{E}[\Delta E_i \mid \Delta H_i = h] &= \mathbb{E}\left[\mathbf{1}\{Z_i + \lambda h \geq 0\} - \mathbf{1}\{Z_i \geq 0\} \mid \Delta H_i = h\right] \\
&= P(Z_i + \lambda h \geq 0 \mid \Delta H_i = h) - P(Z_i \geq 0 \mid \Delta H_i = h) \\
&= P(Z_i \geq -\lambda h \mid \Delta H_i = h) - P(Z_i \geq 0 \mid \Delta H_i = h).
\end{aligned}$$

Let $F_{Z|\Delta H=h}(z) \equiv P(Z_i \leq z \mid \Delta H_i = h)$ denote the conditional CDF of Z_i. Assuming the conditional distribution of Z_i is continuous,

$$\begin{aligned}
\mathbb{E}[\Delta E_i \mid \Delta H_i = h] &= \left[1 - F_{Z|\Delta H=h}(-\lambda h)\right] - \left[1 - F_{Z|\Delta H=h}(0)\right] \\
&= F_{Z|\Delta H=h}(0) - F_{Z|\Delta H=h}(-\lambda h).
\end{aligned}$$

Taking a first-order Taylor approximation of the second term around zero:

$$F_{Z|\Delta H=h}(-\lambda h) \approx F_{Z|\Delta H=h}(0) - \lambda h f_{Z|\Delta H=h}(0),$$

where $f_{Z|\Delta H=h}$ is the corresponding conditional density function. Now the two instances of the term $F_{Z|\Delta H=h}(0)$ cancel out. Hence,

$$\mathrm{E}[\Delta E_i \mid \Delta H_i = h] \approx \lambda h f_{Z|\Delta H=h}(0).$$

Averaging over the distribution of ΔH_i and applying the law of iterated expectations,

$$\begin{aligned}
\Delta p \equiv \mathrm{E}[\Delta E_i] &= \mathrm{E}[\mathrm{E}[\Delta E_i \mid \Delta H_i]] \\
&\approx \lambda \, \mathrm{E}\left[\Delta H_i f_{Z|\Delta H=\Delta H_i}(0)\right].
\end{aligned}$$

Defining $m_i \equiv f_{Z|\Delta H=\Delta H_i}(0)$ (density/slope at the employment margin) gives

$$\Delta p \approx \lambda \, \mathrm{E}[m_i \Delta H_i].$$

Then, using

$$\Delta H_i = \theta_S \Delta S_i + \Delta g_i,$$

we obtain

$$\begin{aligned}\Delta p &\approx \lambda \,\mathrm{E}[m_i \Delta H_i] \\ &= \lambda \,\mathrm{E}[m_i(\theta_S \Delta S_i + \Delta g_i)] \\ &= \lambda\theta_S \,\mathrm{E}[m_i \Delta S_i] + \lambda \,\mathrm{E}[m_i \Delta g_i].\end{aligned}$$

Next, I assume that m_i is independent of ΔH_i. This assumption simplifies the formula. In practical terms it means that the size of the treatment-induced change in human capital is unrelated to how close an individual is to the employment margin.

$$\mathrm{E}[m_i \Delta H_i] = \mathrm{E}[m_i]\mathrm{E}[\Delta H_i].$$

Define

$$a \equiv \lambda \,\mathrm{E}[m_i] \geq 0.$$

For exposition, let

$$\Delta \bar{S} \equiv \mathrm{E}[\Delta S_i], \qquad \Delta \bar{g} \equiv \mathrm{E}[\Delta g_i].$$

Since

$$\mathrm{E}[\Delta H_i] = \theta_S \,\mathrm{E}[\Delta S_i] + \mathrm{E}[\Delta g_i],$$

we obtain

$$\boxed{\Delta p \approx a\theta_S \Delta \bar{S} + a\Delta \bar{g}}.$$

A.6 Methodological Details

A.6.1 Estimation of effects on parents using differences-in-differences/event study

Since I can observe the labor market outcomes of parents before and after the children apply for youth rehabilitation allowance, I can use the panel data in a differences-in-differences or event study design. I will compare the parents approved and rejected applicants prior and after the application.

$$Y_{it} = \alpha_i + \sum_{j=-5, j\neq -2}^{7} \beta_j \mathbf{1}\{\text{Time} = j\} + \gamma_t + \varepsilon \tag{15}$$

where α_i are the individual fixed effects and γ_t include time, year, and parent age fixed effects. The coefficient of interest is then the vector of β_j's, the difference between the approved and rejected applicants in period j, compared to the difference in the baseline period $j = -2$. Since the specification compares only the approved and rejected applicants, I avoid the issues raised in the newest differences-in-differences literature.

A.6.2 Selection-on-observables approach

I implement the covariate-adjustment design using the interactive DDML estimator described in Section 4.3. For each outcome, the procedure estimates two nuisance functions: the conditional outcome regression, $E[Y \mid X, D]$, and the propensity score, $E[D \mid X]$, where D indicates approval. The outcome regression short-stacks predictions from OLS and gradient-boosted trees, while the propensity-score model short-stacks predictions from logistic regression and gradient-boosted trees. The DDML estimate is then formed from the cross-fitted nuisance predictions.

I use five cross-fitting folds. Fold assignment is clustered by applicant, so all observations for the same individual remain in the same fold. This prevents another observation for an applicant from entering the training sample used to predict that applicant's outcome or treatment status. Inference is also clustered at the applicant level. I estimate the nuisance functions separately within each analysis sample, including the full applicant sample and the labor-market-prospect subgroups, rather than imposing nuisance relationships estimated in one population on another.

The specification combines transparent linear learners with a gradient-boosted tree learner that can capture nonlinearities and interactions. Cross-fitting and the orthogonal DDML score reduce the first-order sensitivity of the treatment-effect estimate to errors in these nuisance models.

For the coefficient-stability diagnostic reported with the DDML estimates, I first estimate an OLS model containing the same covariate set and cluster its standard errors by applicant. I calculate the Oster δ for a zero treatment effect using $R_{\max} = \min\{1, 1.3R^2\}$, where R^2 is from the controlled regression. While the OLS-based specification does not perfectly correspond to the DDML specification, a benefit is that the Oster δ is a familiar statistic that can be compared with other settings, unlike the approach of Chernozhukov et al. (2026) designed for these cases.

A.6.3 Overlap and placebo diagnostics for the high-DI-risk subgroup

The main text reports the corresponding diagnostics for the full applicant sample. Figure A.20 shows the estimated propensity-score distributions for approved and rejected applicants in the high-DI-risk subgroup. Figure A.21 applies the DDML procedure to pretreatment outcomes and fixed characteristics that approval cannot affect. Together, the figures assess whether approved and rejected high-risk applicants have common support and whether the adjustment procedure removes the large raw differences between them along observed pretreatment dimensions.

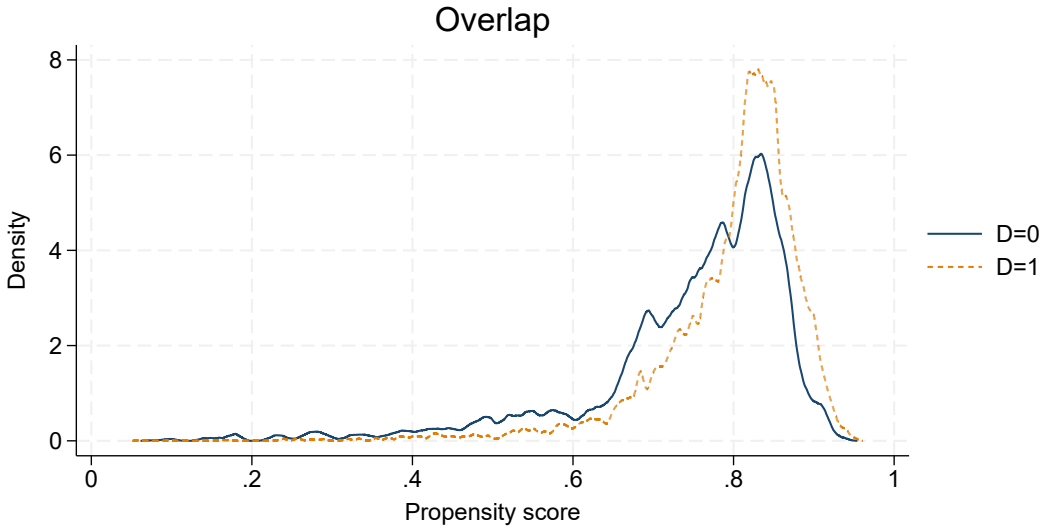


Figure A.20: Overlap for the covariate-adjustment design in the high-DI-risk subgroup.

Notes: The figure shows the distributions of the propensity scores estimated by the DDML procedure using the full covariate set. $D = 1$ denotes approved applicants and $D = 0$ rejected applicants. The propensity score is estimated for the sample observed in years 6–11 after the first application; its distribution may therefore differ in other analysis samples. Figure 6 reports the corresponding figure for the full applicant sample.

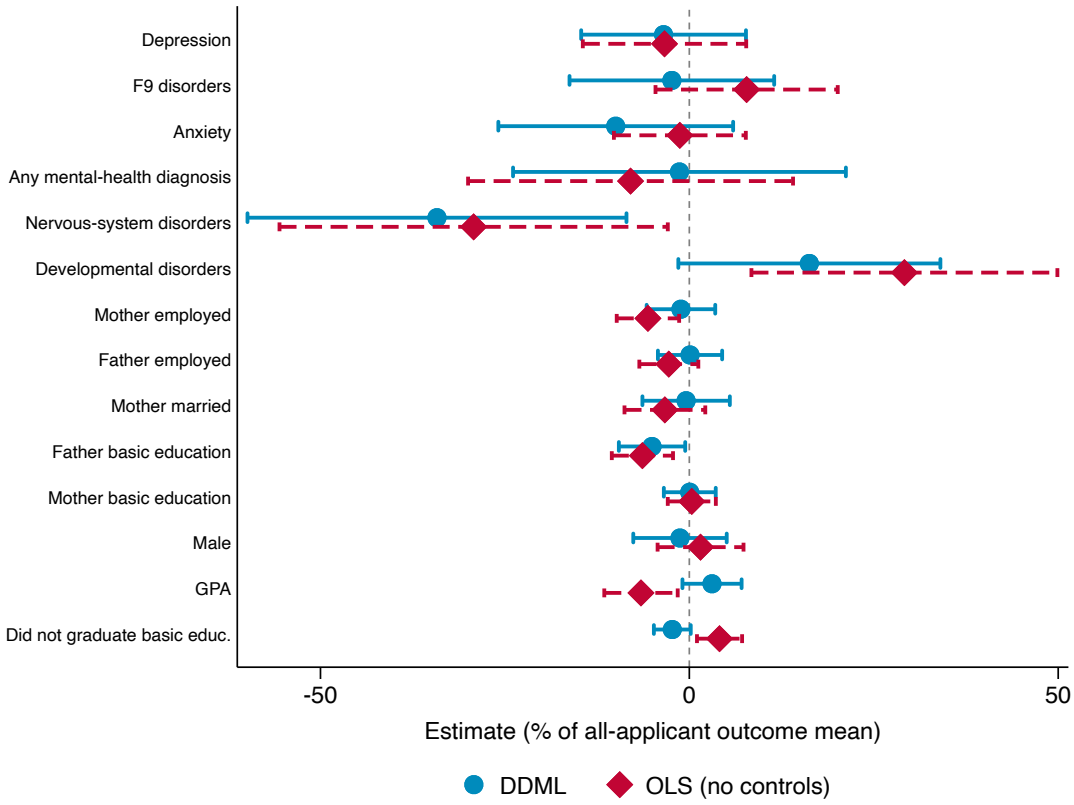
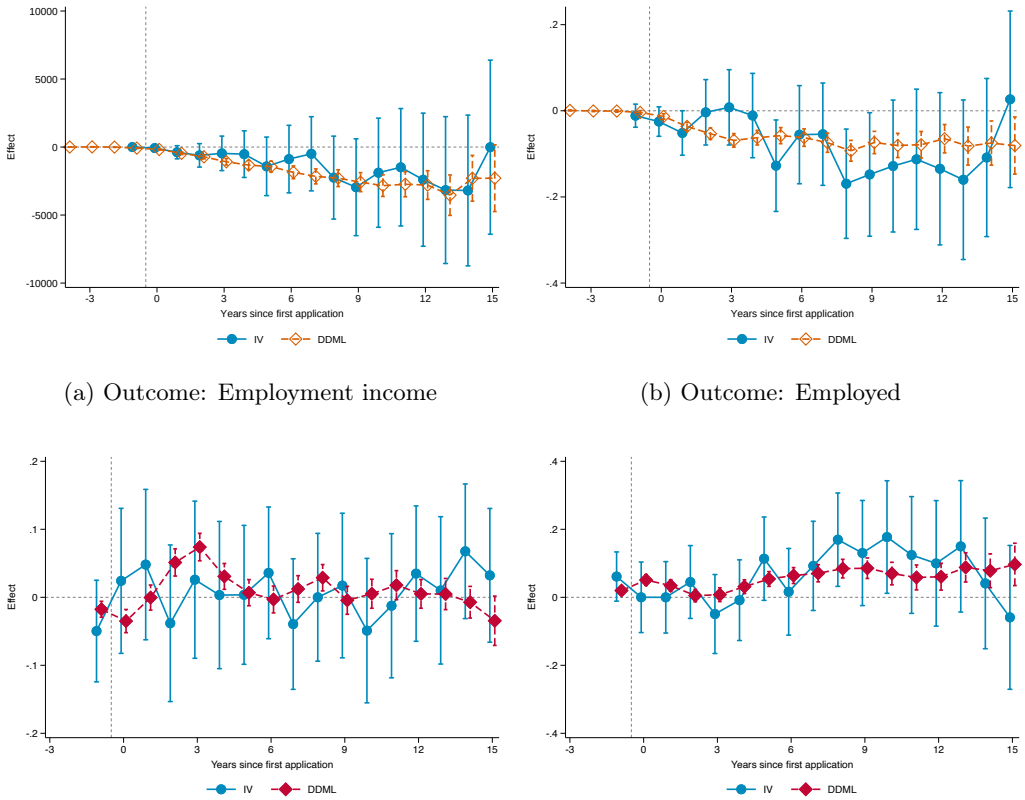


Figure A.21: Placebo and balance tests for the covariate-adjustment design in the high-DI-risk subgroup.

Notes: Each DDML estimate uses the characteristic listed in the figure, measured before the application, as the outcome. Closely related variables and variables constructed from the outcome are omitted from that regression's covariate set. Figure 7 reports the corresponding figure for the full applicant sample.

A.6.4 Additional DDML-IV comparisons



(a) Outcome: Employment income

(b) Outcome: Employed

(c) Outcome: Student

(d) Outcome: NEET

Figure A.22: Complier-weighted DDML and IV estimates.

A.6.5 Cost-effectiveness and MVPF calculations

This section presents some details of the calculation of costs and benefits omitted from the main body of the paper. These provide the arguments for the choices I have made. Robustness to all alternative choices is hard to ascertain, as the number of different decisions that could be made is rather large.

The estimates are produced with similar methods to the rest of the paper, looking at only the sample that can be followed until age 27. I calculate cumulative amounts using only post-application data, expressed in 2023 euros and discounted at an annual rate of 3%. The IV specification is the one in the IV method section, while the DDML choices and sample definition are explained in the corresponding sections.

Most of these numbers include considerable uncertainty and should be taken as suggestive. Since administrative data sometimes include outliers that have very large costs associated with them, I winsorize all costs at the 95th percentile.

***Taxes*:** I use the amount of taxes paid from the FOLK TULO module, adjusted for inflation. This includes the taxes paid on the Youth Rehabilitation Allowance itself.

***Other transfers*:** I similarly use the amount of transfers received from the FOLK TULO module, adjusted for inflation. This includes all transfers.

***Criminal Justice costs*:** I use an estimate of 495 euros per police investigation (Hinkkanen, 2014; Valtioneuvosto, 2026). This is reported as cost per offence. In the data, one offence can have multiple reports associated with it. To get closer to the real costs, I cap the number of offences to 10 per month. While somewhat arbitrary, this helps to remove the possibility that single crimes entered into the system in erratic fashion would drive the results. In addition, for convictions I use a cost of 2000 euros as estimated in Valtioneuvosto (2024). This cost does not include costs of providing punishment, like costs of prison time. As young individuals are not commonly convicted to a closed prison but instead receive suspended sentences, there are very few that receive prison time in my sample. I do not attempt to estimate the costs of providing punishment.

***Victimization costs*:** I value the harm to victims at 2,000 euros per property offence and 10,000 euros per violent or sexual offence. These are round numbers based on the unit costs of crime for England and Wales in Heeks et al. (2018). I use only their costs as a consequence of crime (property stolen or damaged, physical and emotional harm, lost output, and health and victim services) and leave out their police and criminal justice costs, which are already counted under criminal justice costs above. In 2015/16 prices, these consequence costs range from £580 (theft from a vehicle) to £4,670 (theft of a vehicle) for property offences, and from £3,750 (violence without injury) to £31,450 (rape) for violent and sexual offences. Unlike the other items, these costs fall on the victims rather than the government, so they enter willingness to pay and not the net cost.

***Costs of providing education*:** It would be misleading to use the average educational costs since many of the costs of education are fixed (buildings etc.). I instead use

as an estimate of the marginal cost the average cost of materials and supplies in general upper secondary schools, which is estimated to be 476 euros per student-year (Opetushallitus, 2025). I do not value the private costs of education in this calculation.

***Healthcare costs*:** The healthcare costs are based on the healthcare records with information on the provider costs. This information is not available for all of the records. For instance, prior to year 2012 this information is not recorded, and after this it is also not recorded in some administrative units. I make some steps to arrive at a final number. I calculate the healthcare costs in the following way:
1) Cap the visits to 1 visit per person per day per diagnosis to avoid transfers across administrative units counting as multiple visits.
2) Based on the records with costs indicated, I calculate the average costs per length of visit with a particular main diagnosis in ICD-10 classification.
3) Impute the costs based on the main diagnosis and length of stay for all the visits.
4) Validate that the imputation yields sensible costs for a sample of visits where the costs are known.

These costs are not exactly the same as the calculated costs using the NordDRG weights. My data does not include this information, and it would not even be calculable for all of the years in the data.

I do not include the costs of providing primary care in my estimates. These data are only available from the year 2011 onward, and the data do not include costs nor is there a clear template from the official sources to estimate this with the available information on the nature and length of each visit. The point estimates suggest small effects, so this omission is unlikely to turn around the conclusions of the MVPF calculation.

***Medical drug costs*:** The Kela Drug expenses data include the public sector costs of the drugs for each purchase, which I use directly.

A.6.6 Empirical Bayes shrinkage and simulation evidence

Following the teacher value-added literature (Chetty et al., 2014), I shrink examiner leniency toward the unweighted mean across examiners. The starting point is residualized leave-one-out approval:

$$z_{ij} = \frac{1}{N_j - 1} \times \left(\sum_{k=1}^{N_j} (Approved^*_{jk}) - Approved^*_i \right) \tag{16}$$

where $Approved^*$ is the residual from regressing approval on region × year × language fixed effects. The shrunk instrument is

$$z_{ij}^{EB} = \left(\frac{\sigma_z^2}{\sigma_z^2 + s_j^2} \right) z_{ij} + \left(\frac{s_j^2}{\sigma_z^2 + s_j^2} \right) \bar{z} \tag{17}$$

For examiner j, let $\bar{r}_j = N_j^{-1} \sum_{k=1}^{N_j} Approved^*_{jk}$ denote the full-caseload residual mean. I calculate

$$\hat{\sigma}_j^2 = \frac{1}{N_j - 1} \sum_{k=1}^{N_j} (Approved^*_{jk} - \bar{r}_j)^2, \qquad s_j^2 = \frac{\hat{\sigma}_j^2}{N_j}.$$

With J examiners, the shrinkage target and estimated signal variance are

$$\bar{z} = \frac{1}{J} \sum_{j=1}^{J} \bar{r}_j, \qquad \sigma_z^2 = \max \left\{ \frac{1}{J-1} \sum_{j=1}^{J} (\bar{r}_j - \bar{z})^2 - \frac{1}{J} \sum_{j=1}^{J} s_j^2,\ 0 \right\}.$$

Each examiner receives equal weight in these two calculations. The implementation uses the full-caseload sampling variance in the shrinkage weight and applies that weight to the leave-one-out mean; it does not substitute $\hat{\sigma}_j^2/(N_j - 1)$ for s_j^2. The shrinkage step imposes no additional minimum-caseload cutoff beyond the sample used to construct leniency. The within-examiner variance and leave-one-out mean require at least two observations per examiner.

I show using simulations that IV using empirical Bayes shrinkage appears to be unbiased and the standard deviation of the estimates is smaller than with the standard leave-out procedure. In these simulations I use 200 judges, each handling between 10 and 50 cases. Treatment is determined by judge leniency, randomly drawn from uniform distribution, confounder and a random component. A continuous outcome variable is determined by a random component, confounder, and possible treatment. I estimate treatment effects using OLS, 2SLS (not JIVE/leave-one-out), oracle type 2SLS using true leniency (serves as a comparison), standard leave-one-out, and my EB procedure. I repeat this procedure 1000 times to get a distribution of the estimates.

As can be seen in Fig. A.23, the (in practice unattainable) true leniency performs best. Shrinking the estimated leniencies toward the mean of all judges does improve from the standard leave-one-out procedure: the estimates are more centered around 0. In Fig. A.24 it can be seen that when the instrument(s) is/are weaker, the performance of EB in comparison to standard leave-one-out procedure is improved even more.

In practice, after shrinking, I use the shrunk instrument just as any other instrument,

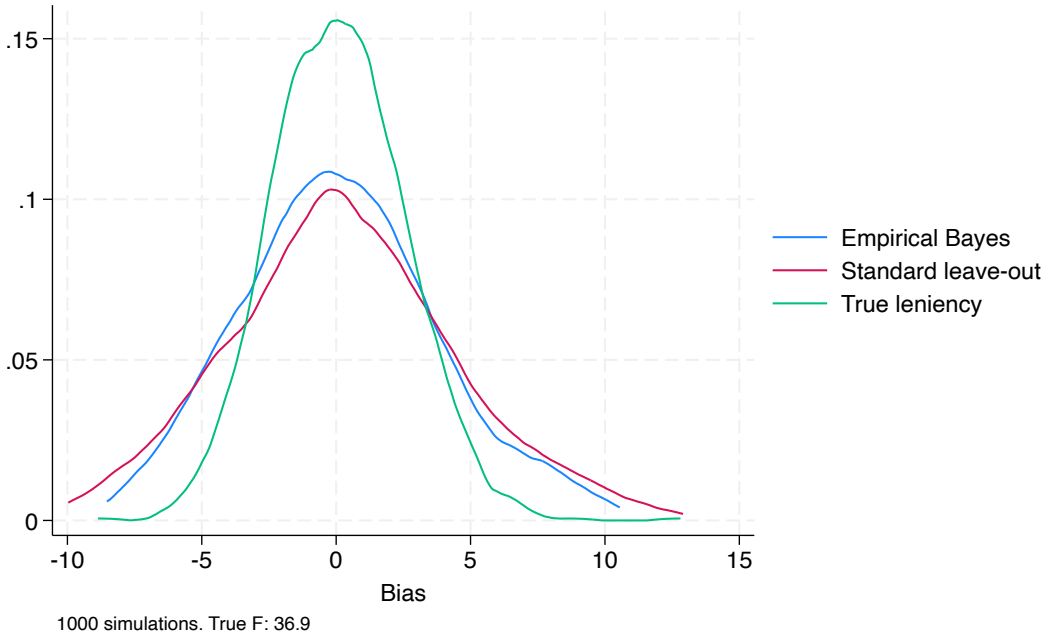


Figure A.23: Distribution of the biases from the second-stage estimates.

Notes: 1000 simulations using different estimators.

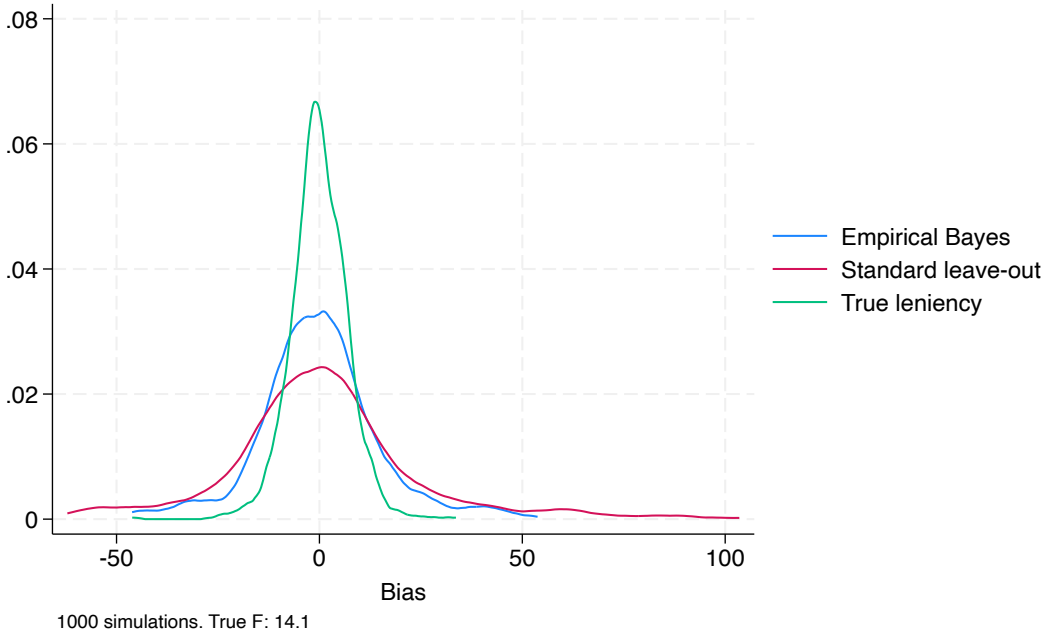


Figure A.24: Distribution of the biases in the second-stage estimates.

Notes: 1000 simulations using different estimators. With lower "true F" (F-statistic if the true leniency is known), the EB-IV substantially improves from the standard leave-out/jackknife estimator.

with no special steps in inference to account for the fact that the EB-IV is estimated. This follows the previous literature on Empirical Bayes. In my simulations I find that the SEs provided by the JIVE procedure are on the conservative side, while the SEs from the EB-IV have roughly correct coverage, though this might not be true with all configurations of instrument strength and other factors.

Empirical Bayes shrinkage in examiner designs. Empirical Bayes shrinkage has previously been used to reduce noise in examiner-specific propensities. Humphries et al. (2025) use Beta-Binomial empirical Bayes shrinkage of leave-one-out judge propensities as a robustness check, shrinking estimates toward either the overall mean or a district-year-specific mean. My implementation uses the same basic idea, but applies a linear empirical Bayes shrinkage rule to the residualized leave-one-out approval rates, with the amount of shrinkage determined by the sampling variance of each examiner's estimated leniency.

More recently, Frandsen et al. (2026) study this idea formally. They show that shrinking noisy leave-one-out first-stage propensities can increase the precision of examiner-IV estimates without additional identifying assumptions or changing the underlying causal estimand and derive the shrinkage rule that minimizes the asymptotic variance of the IV estimator. Their optimal rule additionally depends on first- and second-stage error moments, whereas I use a simpler approach. The simulations in this section examine the finite-sample performance of this implementation.