Shape Up or Ship Out: Causal Effects of Performance Standards in Higher Education

Sander de Vries*

March 26, 2026

Abstract

Many higher education programs dismiss students who perform below some academic standard. Policymakers justify these rules as a way to incentivize higher performance and redirect low-performing students toward more suitable career paths. Exploiting the rollout of performance standards across Dutch bachelor programs, I show that these objectives are largely not achieved. Although the standards create strong incentives and redirect many low-performing students toward alternative programs, they slightly reduce degree attainment, do not shorten time in education, and have no measurable earnings effect. Survey evidence suggests that performance standards decrease student utility. The results question the effectiveness of performance-based dismissal policies.

Keywords: higher education, performance standards, dismissal policy, major choice

JEL Codes: I21, I23, I28

*Department of Economics, Vrije Universiteit Amsterdam, s.de.vries@vu.nl. I gratefully acknowledge valuable comments from Nadine Ketel, Maarten Lindeboom, Hessel Oosterbeek, Stephen Machin, Dinand Webbink, Genia Rachkovski, Julius Ilciukas, Stanislav Avdeev, and conference and seminar participants in Amsterdam, The Hague, Rotterdam, Leiden, Trier, London, Munich, and Catanzaro. The non-public microdata used in this paper are available via remote access to the Microdata services of Statistics Netherlands (project agreement 7930).

1 Introduction

Only two in five students in OECD countries who begin a bachelor's program complete it on time, while the rest either drop out or graduate with delay (OECD (2025)). This imposes substantial costs on both students and governments. For students, each additional year in education delays labor market entry and raises tuition expenses, while leaving without a degree can reduce lifetime earnings (Ost et al. (2018)). Governments bear much of the costs through subsidized tuition and lose potential tax revenue when students spend fewer years in the labor market. Across OECD countries, public and private spending on tertiary education averages about $23,000 per student per year (OECD (2025)). Extended or incomplete studies thus create a substantial fiscal burden, even before accounting for lost output.

To limit these costs, many institutions worldwide use performance standards that dismiss students who perform below minimum academic requirements.[1] Policymakers view such performance standards as tools to raise graduation rates and reduce time in education through two main channels. First, they may increase academic achievement by attaching a penalty to low performance. Second, they may prevent late dropout or prolonged completion times by redirecting students unlikely to succeed toward alternative career paths at an early stage. Despite their widespread use, evidence on the effectiveness of performance standards is limited. Prior work has focused on students narrowly passing or failing performance standards, but shed little light on how students would perform in the absence of a performance standard. It therefore remains unclear whether the presence of performance standards improves performance, matching, and ultimately aggregate education or labor market outcomes.

This is the first study to provide a comprehensive evaluation of the effects of implementing performance standards. I exploit the staggered introduction of performance standards in 351 university bachelor programs in the Netherlands between 1994 and 2014. These performance

[1] For instance, many higher education institutions worldwide dismiss students who score below certain GPA thresholds (e.g., US, Canada, Australia, New Zealand, Mexico, South Korea, Singapore, India), fail the same course multiple times (e.g., Germany, Switzerland, Denmark, Russia, and Austria), or fail a specific number of courses (e.g., the Netherlands, Indonesia, and South Africa).

standards require students to pass a specific number of courses by the end of their first year, and failure results in removal from the program. The requirements are strong: on average, one in five students are dismissed.[2] This means that the risk of dismissal extends beyond the very lowest ability students. Combining a difference-in-differences design with administrative data covering over 700,000 students allows me to study the causal effect of these strict performance standards on both short- and long-term outcomes.

This approach identifies the average effect of introducing performance standards on all students who initially enroll in treated programs. Unlike approaches that focus on students at the margin of failing, the estimates capture the combined effects on all potentially affected students: lower-ability students, who may exert more effort to remain enrolled or switch to more suitable career paths when dismissed, and higher-ability students, whose peer environment and effective class size change as weaker students exit. The results thus speak directly to the policy-relevant question of whether performance standards measurably improve aggregate educational and labor market outcomes.

I find that the performance standards substantially alter student progression in the initial program. They increase dropout after the first year by 32 percent, which implies that 7.5 percent of students are dismissed who would continue otherwise.[3] The completion rate of all initially enrolled students decreases by 3.6 percentage points, which suggests that at least half (3.6/7.5) of the additional dropouts would graduate in their initial program in the absence of the policy. The average enrollment duration of all initially enrolled students falls by 3 months, while the average completion time of graduates is unaffected. This indicates that the shorter enrollments are driven by earlier exit of dismissed students, rather than faster progress of non-dismissed students.

The performance standards do not improve long-run outcomes. Although their goal is to

[2] This is estimated by the Dutch Ministry of Education through a survey (Ministerie van Onderwijs, Cultuur en Wetenschap (2010)). As explained in greater detail in section 3, I only observe dropout status, which can be voluntary or due to dismissal.

[3] The total share of students with insufficient credits is about 20 percent. This implies that over half of dismissed students (12.5/20) would also drop out in the absence of the standard.

improve graduation, I find that instead performance standards slightly reduce average degree attainment.[4] Moreover, they do not reduce total time in higher education, another primary policy goal, nor do they affect employment or earnings up to twelve years after enrollment. The estimates are precise and rule out economically meaningful effects with high confidence. I also find no evidence of positive effects in specific fields or cohorts, programs with many late dropouts, or for students with high ex-ante risk of failure.

A primary threat to the research design is that prospective students avoid programs with performance standards. I find no evidence of such deterrence. Enrollment numbers remain unchanged on average and across adoption cohorts or fields of study. The composition of new students by gender, socioeconomic background, and ability is also unaffected by the introduction of performance standards. These findings suggest that my results are unlikely driven by differential selection of incoming students.

Finally, I explore unintended negative consequences of performance standards. Beyond administrative burdens for program administrators, they may reduce student welfare by increasing pressure to perform or by removing students' potentially preferred option to continue in their initial program. To gauge this disutility, I presented 333 first-year students with hypothetical choice scenarios that trade off monetary gifts against immediate removal of the policy. More than 80 percent were willing to forgo substantial gifts, with 26 percent willing to forgo over €1,700.[5] This suggests that first-year students experience disutility from performance standards. Using administrative data, I also test whether the policy increased antidepressant prescriptions, but find no evidence in support of this.

The results raise concerns about the effectiveness of similar performance-based dismissal policies in less forgiving contexts than the Netherlands. Dutch higher education is characterized by relatively low tuition fees, minimal selective admission, and a high density of

[4] The Association of Dutch Universities states that the objectives of the performance standards are to 'reduce the high dropout rate in bachelor education, shorten the average study duration, and help students find a suitable program as quickly as possible' (Universiteiten van Nederland (2023)).

[5] For reference, the average monthly income of students is estimated at €943 (Groen and Houtsma (2021)) and the annual tuition fee is €2,530.

similarly ranked institutions. These conditions enable most dismissed students to re-enroll in alternative programs. In contrast, in countries with higher tuition, stricter admission, or fewer nearby options, dismissed students may be more likely to leave higher education entirely. Consistent with this, Ost et al. (2018) show that in the United States, students who narrowly fail a dismissal threshold often exit college and experience large earnings losses. In such contexts, performance gains among non-dismissed students would have to be substantial to offset the adverse consequences for those who are dismissed.

This paper adds to several strands of literature. Most closely related are studies on the effects of academic probation, which issues a formal warning to low-performing students and leads to dismissal if performance does not improve. Prior work finds that academic probation improves short-run academic performance for students near the probation threshold, but also increases dropout and induces shifts toward easier courses (Lindo et al. (2010), Fletcher and Tokmouline (2018), Casey et al. (2018), Wright (2020)).[6] While these studies offer valuable insights, they leave two key questions for policymakers: (i) how would students perform in the absence of any performance standard, and (ii) how do performance standards affect the average rather than the marginal student.[7] These questions matter because the presence of performance standards may alter effort and performance, and because effects on low- or high-performing students likely differ from those on students near the threshold. This paper addresses these gaps by testing, for the first time, whether the presence of performance standards improves average educational or labor market outcomes of all affected students.[8]

The findings also contribute to the literature on the optimal design of performance in-

[6] Relatedly, Albert and Wozny (2024) study effects of more intensive academic probation which includes mandatory study time, and Canaan et al. (2023) and Ost et al. (2018) study effects of group coaching and academic dismissal for students already on probation.

[7] To my knowledge, only two papers have considered these aspects before, and they also focus on performance standards in the Netherlands (Arnold (2015), Sneyers and De Witte (2017)). Both consider only the share of students graduating within four years. Using more detailed administrative data, this paper examines whether total graduation rates, study duration, or labor market outcomes are affected.

[8] This distinction is similar to the literature on high-stakes exit exams, where some papers use regression discontinuity designs to study the effects of failing exit exams, and others use difference-in-differences designs to study the effects of the presence of (high-stakes) exit exams (e.g., Ou (2010), Clark and Martorell (2014), Caves and Balestra (2018), ter Meulen (2023), Fidjeland (2023)). A similar distinction arises in the literature on school accountability systems (Figlio and Loeb (2011)).

centives in education more generally. Previous papers often consider relatively small-scale interventions, short-run outcomes, and (for ethical reasons) positive incentives (e.g., Angrist and Lavy (2009), Angrist et al. (2009), Leuven et al. (2010), Leuven et al. (2011), Rodríguez-Planas (2012), Bursztyn and Jensen (2015), Lavecchia et al. (2016), Papay et al. (2016), Clark et al. (2020), Hvidman and Sievertsen (2021)). However, results from small-scale interventions can differ greatly from scaled-up policy implementations (Al-Ubaydli et al. (2017)), negative incentives can have different effects than positive incentives due to asymmetric behavioral responses (Levitt et al. (2016)), and short- and long-run effects of interventions targeting human capital are often different (Almond et al. (2018), Aizer et al. (2024)). This paper contributes by providing unique evidence on the *long-run* effects of a *large-scale* implementation of a *negative* performance incentive.

This paper proceeds as follows. I provide the institutional context in section 2. Section 3 presents the data and descriptive statistics. I discuss identification in section 4 and present the main results in section 5. Section 6 explores effects on student wellbeing. Section 7 concludes.

2 Institutional context

2.1 Higher Education in the Netherlands

Higher education in the Netherlands is offered by two types of institutions: research universities and universities of applied sciences. This paper focuses exclusively on bachelor programs at all 13 research universities. Bachelor programs offer specialized undergraduate education throughout the entire program, unlike, for example, in the United States, where students specialize later. All programs have a fixed curriculum in the first year, so strategic course taking in response to performance standards is not possible. Eligibility for these programs requires a high school diploma from the preparatory academic track or completion of the first-year curriculum at a university of applied sciences.[9]

Dutch university education possesses two distinctive features relevant to this paper. First, the majority of bachelor programs are required to admit all eligible applicants. Only a small number of programs are oversubscribed, in which case students are selected based on lotteries or additional information from CVs, interviews, and motivation letters. Hence, conditional on eligibility, admission is largely non-selective, and students who drop out are generally free to enroll in other programs. Second, universities in the Netherlands receive public funding, staff salaries are determined through collective agreements, and tuition fees for bachelor programs are standardized by law. As a result of the broad accessibility for students across programs, and the uniformity in terms of university funding and tuition fees, Dutch universities are considered to be qualitatively very similar.[10]

The bachelor programs use the European Credit Transfer System (ECTS).[11] The programs consist of a combination of courses and a thesis, amounting to a total of 180 credits, which are awarded upon passing courses or the thesis. Ideally, students are expected to achieve 60 credits each year. However, in practice, students have the flexibility to adjust their course load or they may fail certain courses, causing delays in their study completion. Apart from the performance standards and personal costs and benefits, students faced few incentives to perform well or graduate on time. The main external incentive during the study period was that most students received a government grant for higher education that required graduation from any program within ten years.[12]

[9] For reference, in 2018, 19 percent of secondary school students in the Netherlands completed this track. Students from international backgrounds must present qualifications equivalent to the Dutch preparatory academic track to gain admission. Admission to some STEM programs further requires completion of designated STEM courses in secondary school.

[10] To support this point empirically, Avdeev et al. (2024) show that the standard deviation of the ranks of Dutch universities in the Times Higher Education ranking from 2023 is much lower than that of universities in the US, Sweden, Croatia, and Chile.

[11] University programs have been split into bachelor and master programs since the Bologna reform in 2002. Before the reform, programs were undivided, and upon completion of a program, one was granted the equivalent of today's master's degree. The ECTS was also implemented as part of this reform. Before 2002, programs used a similar metric to assign credits to passed courses. While I also include undivided programs in my analysis, 85% of the programs introduced the performance standard after 2003 and therefore my results mostly apply to bachelor programs (rather than undivided programs).

[12] Students who failed to graduate within ten years had to repay the grant. Between 1993 and 1996, the

2.2 The Performance Standard at Dutch Universities

In the early 1990s, universities struggled with high dropout rates and long study durations, and this was particularly salient among students who performed poorly in their first year. This raised concerns among policymakers about the efficiency of university education. To address these issues, the government introduced a 1993 law requiring universities to advise all first-year students on whether to continue in their program. The government viewed such advice as a promising instrument to redirect students unlikely to succeed toward more suitable fields of study. Although all programs were required to issue an advice, universities retained discretion over whether it was binding.[13] Under a binding regime, known as the Binding Study Advice (BSA), students who receive a negative advice are dismissed from their program. Dismissed students remain eligible to enroll in other programs, either within the same institution or elsewhere. Credits for closely related courses completed in the initial program may be transferred to the new program, but recognition is discretionary and depends on the rules of the receiving program.[14]

When universities implement the BSA, they must adhere to specific requirements. First, they must establish clear criteria for issuing positive advice. This must include a minimum number of credits students must earn by the end of their first year. On average, this threshold requires students to pass 65% of their courses, though there is some variation among programs. Some programs may also have additional requirements, such as mandatory completion of specific courses or policies that allow students to offset failed courses with high grades in others. Second, the advice must be provided at the end of the first year, and programs are required to inform students in advance if they are at risk of receiving negative advice. Finally, universities must consider personal circumstances when making these

grant was conditional on yearly performance.

[13] The estimates therefore capture the effects of having a binding rather than a non-binding study advice. The regulatory framework is specified in the Higher Education and Scientific Research Act (in Dutch: Wet op het Hoger Onderwijs en Wetenschappelijk Onderzoek, WHW) of 1993.

[14] Information on how often programs allow credit transfer is unavailable. Several universities explicitly prohibit the transfer of credits for first year courses. Because most dismissed students in Section 5 enroll in substantively different curricula, credit transfer are likely limited for this group.

decisions, and students have the right to contest negative advice.[15]

3 Data

I use administrative data from Statistics Netherlands covering the universe of students enrolled at Dutch universities. The data link individuals across multiple government registers, including higher education enrollments, graduation records, earnings, and prescription drug use.

Sample. The sample includes all students who first enrolled in a bachelor program at a research university between 1991 and 2014.[16] Each observation in my dataset represents a student and his/her initial enrollment. If a student later switches programs and re-enrolls as a first-year student elsewhere, they are not recorded as a new observation. This approach defines the initial enrollment as the start of the treatment period, treating any subsequent changes as outcomes. I exclude students who were above age 30 at the start of their studies (1.5%) and students in programs that (i) never implemented a performance standard (2.4%), (ii) did not implement a performance standard by 2014 (0.8%), (iii) have a performance standard from the start (0.9%), or (iv) which ceased to exist before 2014 (1.5%). Following these selection criteria, my sample covers 712,384 students across 351 distinct bachelor programs. Whenever I refer to a program, I refer to a specific bachelor's program at a specific institution. Below, I describe the construction of the key variables and present descriptive statistics of all students from the core sample.

Student characteristics. Table 1 shows that students are balanced by gender. Most enroll shortly after secondary school at about age 19. International students and those with

[15] In practice, about 4% of students appeal a negative binding study advice, and roughly 20% of them are successful (Ministerie van Onderwijs, Cultuur en Wetenschap (2010)). In addition, Vooren et al. (2024) show that some students continue their studies despite not meeting the minimum course requirements, suggesting that the rules are waived for some students on the basis of personal circumstances. Official figures for this group are, however, not available.

[16] Graduates from Universities of Applied Sciences often enroll in short pre-master's programs at research universities to qualify for a master's degree. However, the data do not distinguish between bachelor's and pre-master's enrollments. I therefore exclude all students who already hold a bachelor's degree from a University of Applied Sciences to ensure that the sample excludes pre-master students.

prior enrollment at a University of Applied Sciences form 23 percent of the population. Using population and tax registers, I link students to their parents and measure the father's income rank.[17] Since not all students, especially international students, can be linked to their parents, this variable is available for 75 percent of students. Average paternal income is high compared to the national mean. For cohorts starting after 2006, I also observe students' average secondary school grade in the final year. This measure is standardized across all other individuals from the same cohort and in the same level of secondary education.

Education outcomes. I first construct an indicator for first-year dropout, defined as not re-enrolling in the same program after the first year. This is relevant because performance standards determine whether students are allowed to continue after the first year. Table 1 shows that 25 percent of students drop out from the initial program after the first year, which can be voluntarily or through dismissal if the program has a performance standard.

The main long-run outcomes are students' degree attainment and time spent in education. Table 1 shows that only 59 percent of students complete their initial degree. Since 25 percent drop out after the first year, this implies that 16 percent drop out after the second year or later. Because this is a sizable group, I also examine in section 4 whether performance standards affect the timing of dropout. Among all initially enrolled students, 78 percent eventually graduate from a research university, and this share rises to 86 percent when including graduation from a University of Applied Sciences.[18] Students spend on average 3.5 years in their initial program and 6.3 years in higher education.[19]

Earnings. To accurately capture long-run earnings trajectories, I measure earnings 12 years after the initial enrollment. This corresponds to an average age of 31. Earnings

[17] Income is first observed in 2001. For cohorts starting after 2001, income is measured in the year of enrollment; for earlier cohorts, income is measured in 2001. The rank is defined relative to all same-age individuals in the Netherlands with observed income.

[18] I observe enrollments and graduations up to 2022. This allows me to track education for at least nine years after entry for the last cohort (2013), with longer follow-up for earlier ones. Among students who entered in 2012, only 1.9% of students were still enrolled after nine years and did not obtain a degree yet.

[19] Although enrollment durations are reported in years for ease of interpretation, the underlying data are recorded monthly, which allows for fine granularity. The time spent in higher education is counted from the start of the initial enrollment until the end of the last enrollment, including breaks like gap-years. The results are the same when I define the enrollment duration as the sum of the yearly enrollments.

Table 1: Descriptive Statistics

	Cohorts	Mean	SD	N
Characteristics				
Male	All	0.50	0.50	712,384
Age	All	19.52	1.63	712,384
International student	All	0.12	0.32	712,384
Previously enrolled in University of Applied Sciences	All	0.11	0.31	712,384
Income Percentile Father	All	0.75	0.25	537,944
Secondary School GPA	2006-2014	0.12	1.00	712,384
Education outcomes				
Dropout after the first year	All	0.25	0.44	712,384
Complete initial program	All	0.59	0.49	712,384
Obtain research university degree	All	0.78	0.42	712,384
Obtain higher education degree	All	0.86	0.34	712,384
Years enrolled in initial program	All	3.50	2.35	712,384
Years enrolled in higher education	All	6.25	2.74	712,384
Labor market outcomes 12 years post enrollment				
Missing earnings records	1991-2012	0.13	0.42	596,996
Positive earnings (conditional on non-missing)	1991-2012	0.96	0.20	518,281
Annual earnings (conditional on positive earnings)	1991-2012	55,409	36,941	496,333
Program size	All	93.99	120.58	351

Notes: This table presents descriptive statistics for all variables considered in the analysis. ‘Cohorts’ refers to the years for which I observe the corresponding outcome. The remaining columns report the mean and standard deviation of each outcome and the number of observations underlying it.

encompasses income from employment and entrepreneurship, and is based on tax-return statements. I adjust earnings to 2015 prices. Because earnings are observed until 2024, this information is only available for cohorts that enrolled before 2012. I do not observe tax-returns for 13 percent of this sample, which is mostly due to migration of foreign students. Among the group for whom I do have tax returns, a small share (4%) have zero or negative earnings. In the analysis, I show that having missing records or having non-positive earnings is not related to being enrolled in a program with a performance standard, suggesting that selection into observed earnings is unlikely an issue.

Program-level outcomes. In the analysis, I collapse the individual-level data to the program level. Unless stated otherwise, the outcomes Y_{pt} are averaged over all students who enrolled in year t in program p. The total number of year-by-program observations is 7,579, which is less than 351 (the number of programs) × 24 (the number of years) = 8,424 because some programs were established after 1991.

3.1 Implementation of the Performance Standard

I collected data on the introduction of the performance standards in programs from two sources. The Association of Dutch Universities provided me with records of programs between 1995 and 2012. For the years 2012 to 2014, I rely on annual reports from programs, faculties, or universities.

Figure 1 shows that the rate of adoption varied substantially across academic programs. Since 2003, additional programs have adopted a performance standard each year. By 2014, all programs in my sample have adopted the policy. Table A1 further documents the adoption at the institution level. Some universities introduced performance standards simultaneously across all programs, while others delegated the decision to faculties, creating within university variation. In total, there are 40 instances where a university adopted a performance standard for at least one of their programs.

The staggered adoption reflects divergent attitudes toward the policy over time and across institutions. The option to dismiss poor performing students introduced in 1993 sparked significant debate among university and faculty boards, which had the authority to decide whether to adopt the BSA in their programs. Many academics and students strongly opposed dismissing students based solely on their first-year performance.[20] Critics also argued that the BSA would add administrative burdens, including an increase in appeals and procedural complexities. As a result, most universities were reluctant to directly implement the policy. At the same time, the Dutch government pressured universities to shorten the time students spend in programs and improve graduation rates, viewing the BSA as a promising strategy to achieve these national objectives. In 2006, the Minister of Education explicitly encouraged institutions to implement the BSA.[21] This likely contributed to the faster and nearly universal adoption during and shortly after this period.

Table A2 shows that the year of adoption is not related to students' average secondary school grades, program size, graduation rates and field of study. This suggests that adoption timing is not strongly related to key program characteristics. Since programs that adopt performance standards at different times are similar along these dimensions, it is plausible that their outcome trends are also comparable. This assumption is central to the identification strategy, which I discuss in the next section.

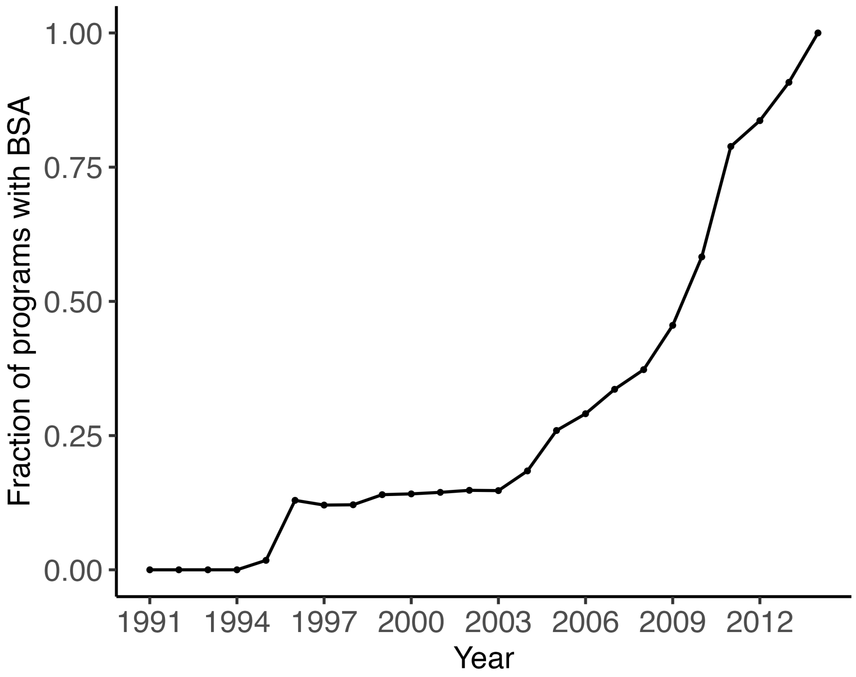
4 Identification

I exploit the staggered rollout of performance standards to estimate their causal effect. This approach compares changes in student outcomes in programs that adopt a performance standard with changes in programs that adopt them later.

[20] For example, the National Students' Union (LSVB) has consistently opposed the BSA since its inception, culminating in protests such as the 2009 occupation of the main building at the University of Groningen (Volkskrant, 2009)

[21] See 'Serieus werk maken van bindend studieadvies', Volkskrant (2006).

Figure 1: Staggered rollout of the performance standard in bachelor programs



Note: This figure presents the share of programs with a BSA in each year between 1991 and 2014. The total number of programs in my sample is 351.

More specifically, I estimate event-study regressions of the form

$$Y_{pt} = \delta + \sum_{l=-5, l\neq -1}^{2} \beta_l I[\tau_{pt} = l] + \alpha_p + \gamma_t + \epsilon_{pt}, \tag{1}$$

where Y_{pt} is the outcome corresponding to program p in year t. Unless otherwise stated, the outcomes are averages taken over all students who are initially enrolled in program p in year t. Moreover, τ_{pt} is the year relative to treatment, with $\tau_{pt} = 0$ in the first treatment period, and α_p and γ_t are program and year fixed effects. The regressions are weighted by the average program size in the pre-treatment periods. In this model, the effects of implementing performance standards l years ago are identified by β_l for $l \geq 0$. The estimates corresponding to $l < 0$ serve as placebo checks.

The key identifying assumption is that, in the absence of a performance standard, the outcomes of students in treated programs would evolve in parallel to that of students in untreated programs. There are multiple factors that support the plausibility of this assumption. First, Dutch bachelor programs are very homogenous. As explained in section 2, except for some oversubscribed programs, all bachelor programs have to admit all students. Moreover, tuition fees are equal, universities are similarly funded, and geographical distances are relatively small. As a result of these features, the composition and quality of bachelor programs is very similar across universities. This makes it more likely that programs experience similar trends. Second, as performance standards are often uniformly implemented for entire faculties or universities, they are unlikely to be implemented in response to undesirable trends in specific programs. This is empirically supported by pre-trend estimates, which show that before implementation treated and untreated programs evolve similarly.

Another key assumption is that programs' outcomes do not depend on the treatment of others. It is unlikely that students in programs with performance standards affect students in other programs directly because programs do not share courses in the first year. However, an indirect violation of this assumption arises when prospective students are deterred by the implementation of performance standards and substitute towards programs without performance standards. This results in a direct effect on the composition of students in treated programs, but it also affects the composition of untreated programs. I test for the importance of such selection effects in multiple ways in subsection 5.1, and find no evidence for compositional changes.

Methodological concerns. Recent papers have shown that standard two-way fixed effects regressions can yield misleading estimates, in particular when the fraction of treated units gets large over time (de Chaisemartin and D'Haultfœuille (2020)). To overcome such issues, I use the Callaway and Sant'Anna (2021) doubly robust estimator. I also use their estimator of the aggregated Average Treatment effects on the Treated (ATT), which is a weighted average of the period-specific ATTs.

Other recently expressed concerns are related to model selection and inference. Raw data rarely exhibits parallel trends for treated and control units, and researchers commonly use different techniques, such as adjusting for covariates, to address this problem. However, conditioning the analysis on passing placebo checks induces pre-testing problems (Roth (2022)), and, more generally, high degrees of freedom in specification choices can result in sizeable replication problems (Menkveld et al. (2024)). To mitigate these concerns, I simply report estimates without conditioning on control variables. As a result, the ATT estimates are simple weighted averages of 2 × 2 difference-in-differences estimates between treated and not-yet-treated programs, with the weights equal to the programs' student shares.[22]

Using this unconditional specification, I find at most mild deviations in pre-trends, and only for a limited number of variables. To check whether the estimates change when more carefully chosen control groups are used, I use Synthetic Difference in Differences (Arkhangelsky et al. (2021)). Unlike the baseline specification, which compares all treated programs to all not-yet-treated programs, this method matches treated programs to not-yet-treated programs with similar pre-treatment outcome trajectories. This approach automates the process of choosing appropriate control groups for each treated unit, all while retaining statistical guarantees. I find similar results using this approach.

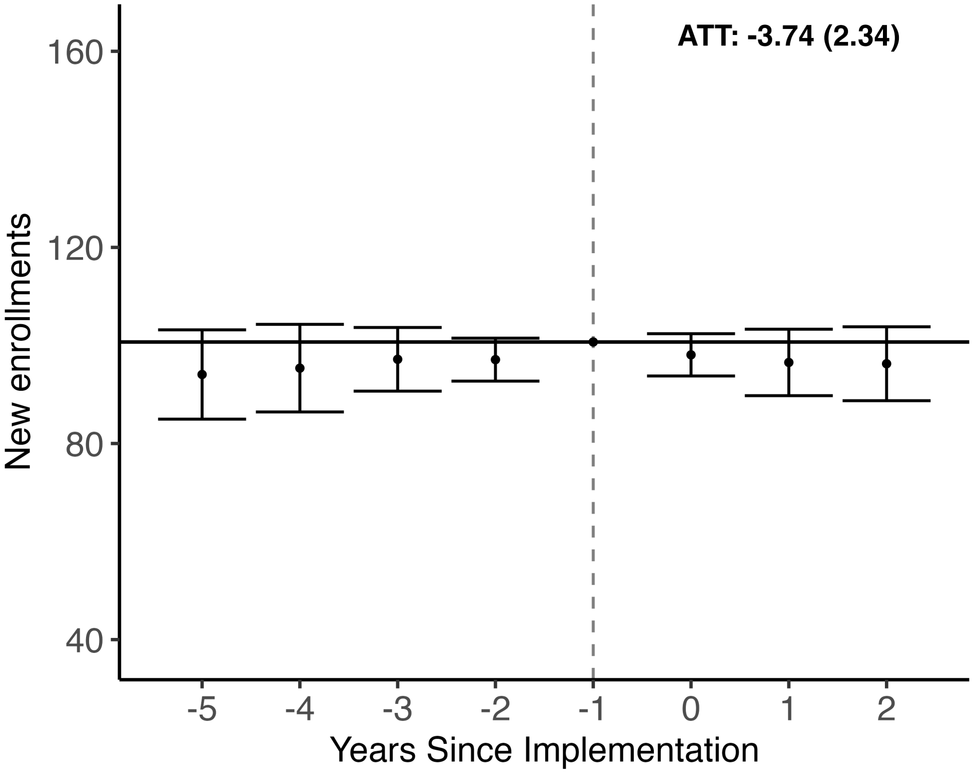
5 Results

5.1 Deterrence effects

In theory, imposing a performance standard is expected to discourage some students from enrolling for two reasons. First, as students cannot perfectly predict their performance in a new program, the performance standard introduces a risk of dismissal. This can deter students who are unsure about their ability to pass the threshold. Second, students who would otherwise exert too little effort to meet the standard are now confronted with the

[22] Moreover, to deal with multiple hypothesis problems in event-study estimates (Freyaldenhoven et al. (2021)), I use simultaneous confidence bands to report confidence intervals.

Figure 2: The effect of performance standards on new enrollments



Note: This figure presents event-study estimates of the effect of implementing performance standards on the number of first-year enrollments. The sample comprises of all 7,579 yearly observations between 1994 and 2014 of the 351 programs. Unlike for the other specifications in this paper, the regression is not weighted by program size, as program size is the outcome of this regression. The ATT is based on the weighted average of the dynamic effects for the first three years after implementation. The graph is centered around the average program size in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing.

decision to raise their effort levels. For some of these students, increasing their efforts may not seem worthwhile, causing them to select another program.

The main outcome where a deterrence effect may be expected to show up is the number of enrollments in the first year. However, Figure 2 shows that imposing a performance standard does not impact the number of new students. The estimate itself is small, and the standard error rules out decreases in program size up to 0.07 standard deviations with 95 percent confidence. With synthetic difference-in-differences estimation, this reduces further to 0.05 standard deviations (Table A5). I find a similarly precisely estimated effect around zero when using the log of program size, indicating that this estimate is not sensitive to functional form

Table 2: The effect of performance standards on the composition of new students

	Income rank father	Male	Age	Previously enrolled at University of Applied Science	Foreign	Secondary school grade	Prob. of graduation within 4 years
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ATT	0.000	-0.006	0.033	-0.002	-0.002	0.033*	0.001
	(0.003)	(0.006)	(0.029)	(0.005)	(0.005)	(0.019)	(0.003)
ATT$_{-1}$	-0.003	0.000	0.010	0.004	-0.005	-0.003	-0.003
	(0.004)	(0.007)	(0.022)	(0.005)	(0.008)	(0.021)	(0.002)
ATT$_{-2}$	-0.001	-0.008	-0.009	0.002	-0.011	0.005	-0.005
	(0.004)	(0.007)	(0.029)	(0.006)	(0.008)	(0.030)	(0.003)
ATT$_{-3}$	-0.002	-0.009	0.014	0.007	-0.014		-0.009*
	(0.005)	(0.008)	(0.032)	(0.006)	(0.011)		(0.005)
ATT$_{-4}$	0.001	-0.008	-0.025	-0.002	-0.019*		-0.012**
	(0.004)	(0.008)	(0.039)	(0.007)	(0.011)		(0.006)
Mean	0.754	0.484	19.430	0.103	0.132	0.110	0.417
SD	(0.041)	(0.232)	(0.490)	(0.086)	(0.145)	(0.403)	(0.175)
N	7,579	7,579	7,579	7,579	7,579	2,792	7,579

Notes: Each column presents estimates of the effect of implementing performance standards on a different outcome. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the dynamic effects for the first three periods after treatment. See Figure A3 for all event-study estimates. The next four rows present placebo estimates based on pre-treatment period differences. The last two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

(Figure A1). Lastly, in Figure A2, I check heterogeneity in effects by studying effects on particular adoption cohorts or specific fields, and find no effects either.

The lack of an effect on program size implies that if a deterrence effect exists, a comparable number of students must be drawn to programs with performance standards to maintain a stable program size. In that case, the composition of programs likely changes. To investigate this, I examine students' socioeconomic status, measured by their father's income, as well as their gender, age, previous enrollment in a University of Applied Sciences, nationality, and average secondary school exam grades. Since grades are available only for the years 2007-2014, I provide pre-trend estimates for up to two years for this variable. Columns (1) to (6) of Table 2 show that these student characteristics do not meaningfully change upon the implementation of performance standards.[23]

I next test whether treated programs attract students with higher predicted probabilities of on-time graduation based on their pre-enrollment characteristics. Specifically, I estimate each new student's likelihood of graduating within four years from their chosen program using the variables described above (excluding grades) and parental education.[24] This approach allows me to test whether students are selectively deterred from programs in which they are expected to perform worse. For instance, while the overall share of male students is unchanged, males may be disproportionately deterred from programs where they perform worse. Consistent with previous results, column (7) shows no evidence that implementing a performance standard increases the ex-ante quality of incoming students.

Finally, in Appendix C, I test for deterrence effects using an alternative identification strategy. I show visually and through regression results that whenever formerly popular programs among individuals who come from the same municipality implement a performance standard, then the number of prospective students from this municipality who enroll in programs with a performance standard increases proportionally. In other words, students from the same municipality consistently choose similar programs over time, and this is unaffected by the introduction of performance standards.

I conclude from this, and the evidence above, that implementing a performance standard does not deter students from enrolling. This also implies that any effects on subsequent outcomes are unlikely explained by a differential selection of students once programs implement performance standards.

The absence of a deterrence effect is surprising, given that roughly 20 percent of students fail to meet the performance standards. Several explanations are possible. First, students may hold strong preferences for their chosen programs, leaving no preferred alternatives even

[23] The only dimension with suggestive evidence of selection is in Column (6), which shows a marginally significant and small increase in secondary school grades ($p = 0.08$). To further assess this, I re-estimate the effect using synthetic difference in differences, and obtain a more precisely estimated zero effect (Table A3).

[24] Details of the prediction exercise are provided in Appendix B.

when it includes the risk of dismissal. Second, students may be overconfident and therefore not perceive or strongly underestimate a risk of dismissal. Third, although programs were by law required to disclose the adoption of performance standards, some students may not have received or understood this information. Since the available data do not allow these mechanisms to be clearly distinguished—and because deterrence effects are not the main focus of this paper—I leave these questions for future work.

5.2 Effects in the initial program

I next estimate effects on students' career trajectories after the first year. At this point, students who did not pass the performance standard are not allowed to continue in their initial program.[25] Figure 3 shows that the introduction of performance standards greatly increases drop-out rates after the first year. The figure is centered around the pre-treatment dropout rate of 23 percent and shows a direct and persistent increase by 7.5 percentage points (32 percent). This indicates that performance standards lead to the dismissal of a significant number of students who would otherwise have continued.[26]

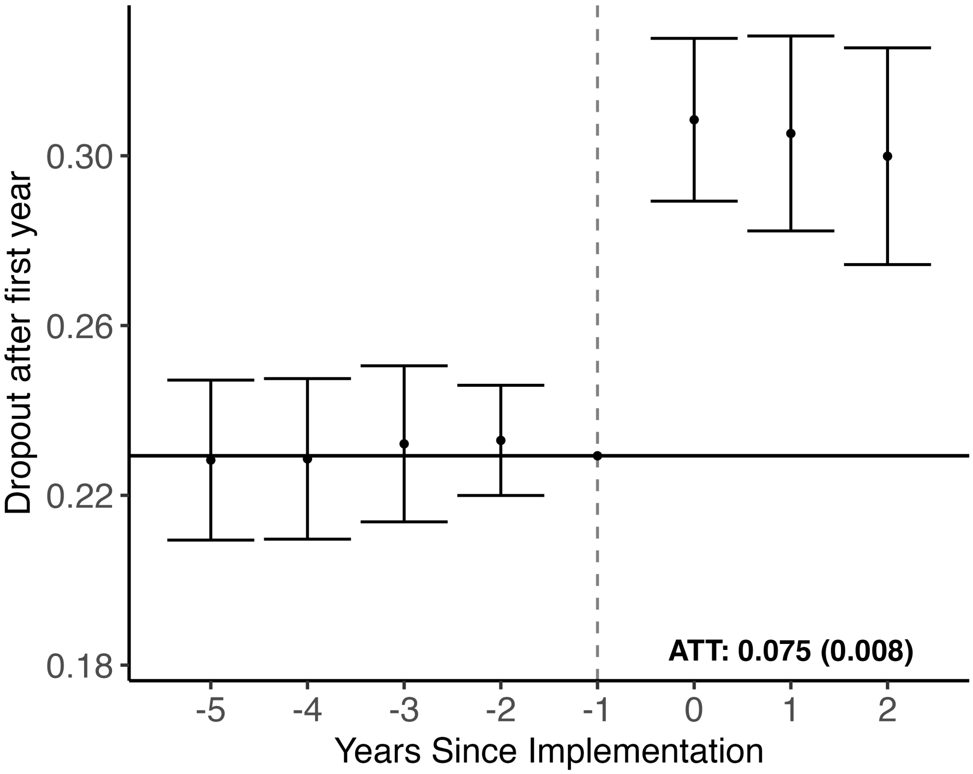
Table 3 shows what students do when they are dismissed. Columns 1 and 2 show that re-enrollment at another institution increases by 3.8 percentage points and dropout from higher education increases by 1.5 percentage points. This suggests that most dismissed students (5.3 of the 7.5 percentage point increase in dropout) make substantial changes in where or whether they study.

Dismissed students also make substantial changes in what they study. Enrollments to the same program at another institution increase by 1 percentage point (column 3), and switching within the same field rises by 1.4 percentage points (column 4). This shows that among the dismissed students that do not leave higher education (6.0 p.p.), the majority

[25] I do not observe the grades or the number of credits in the first year, preventing me to study effects on short-term study performance.

[26] Some dismissed students would also have dropped out in the absence of the performance standard. Given that the total share of dismissed students is estimated to be around 20 percent (Ministerie van Onderwijs, Cultuur en Wetenschap (2010)), this likely holds for over half of the dismissed students.

Figure 3: The effect of performance standards on dropout after the first year



Note: This figure presents event-study estimates of the effect of implementing performance standards on first-year dropout, which is an indicator for whether a student does not continue into the second year of the initially enrolled program. The sample comprises of all 7,579 yearly observations between 1994 and 2014 of the 351 programs. The regression is weighted by the average program size in the pre-treatment periods. The ATT is based on the weighted average of ATT_l for the first three years after implementation. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing.

make substantial changes in what they study: switching to a different field at a research university increases by 2.4 percentage points (column 5), and the remaining 1.2 percentage point enroll at a University of Applied Sciences (column 6).

About half the number of additional dropouts after the first year would have dropped out later. Table 4 shows that performance standards reduce dropout after year two by 2.2 percentage points and dropout after year three or later by 1.8 percentage points. This suggests that performance standards result in the early dismissal of some students who would otherwise have dropped out later in the absence of the policy.

However, the increase in dropout after the first year is larger than the reduction in dropout after the second year and later. As a result, the overall completion rate in the initial program reduces by 3.6 percentage points (column 3). This reduction in the completion rate suggests that about half (3.6/7.5) of dismissed students would have graduated in the absence of a performance standard. This estimate is conservative, because it assumes that the graduation rate of non-dismissed students did not change. If performance standards raised their graduation rates—for example, by improving first-year performance—the implied share of dismissed students who would have graduated is even higher.

One of the primary benefits that programs mention for implementing performance standards is that it reduces the time that students spend in their program. Consistent with this, column 4 shows a reduction of 0.24 years (7 percent, or 0.37 SD) in the time students spend in their initial programs. This decline could arise from three channels: (i) dismissing students who would have dropped out later, (ii) dismissing students who would otherwise have stayed and graduated, or (iii) reducing the enrollment duration of students who always graduate. The last channel would reflect improved study performance among stayers.

However, I find no evidence for this last channel. Column 5 shows a precisely estimated effect close to zero on average completion time. Because this outcome is conditional on graduation, and graduation itself is affected by performance standards, the estimate may reflect selection. Any bias is likely downward, however, because the students who do not

Table 3: The effect of performance standards on education choices after the first year

	Enroll at other institution	Leave higher education	Enroll in same program at research university	Enroll in same field at research university	Enroll in other field at research university	Enroll in University of Applied Sciences
	(1)	(2)	(3)	(4)	(5)	(6)
ATT	0.038***	0.015**	0.010***	0.014***	0.024***	0.012***
	(0.005)	(0.007)	(0.003)	(0.003)	(0.003)	(0.003)
ATT$_{-1}$	0.008*	0.002	0.001	-0.005	0.003	0.002
	(0.005)	(0.003)	(0.001)	(0.004)	(0.004)	(0.004)
ATT$_{-2}$	0.005	-0.004	-0.001	-0.000	0.008	0.000
	(0.005)	(0.003)	(0.001)	(0.005)	(0.006)	(0.004)
ATT$_{-3}$	0.001	0.001	0.001	-0.004	0.006	-0.005
	(0.006)	(0.004)	(0.001)	(0.005)	(0.005)	(0.005)
ATT$_{-4}$	0.000	0.006	0.001	-0.007	0.002	-0.002
	(0.006)	(0.006)	(0.001)	(0.006)	(0.005)	(0.005)
Mean	0.124	0.035	0.005	0.045	0.072	0.072
SD	(0.061)	(0.040)	(0.008)	(0.050)	(0.044)	(0.048)
N	7,579	7,579	7,579	7,579	7,579	7,579

Notes: Each column presents event-study estimates of the effect of implementing performance standards on a different outcome. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the dynamic effects for the first three periods after treatment. See Figure A4 for all event-study estimates. The next four rows present placebo estimates based on pre-treatment period differences. The next two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

graduate because of the performance standards are probably those with longer completion times in its absence. In other words, even though graduates in treated programs are likely positively selected *and* they faced strong incentives to perform during the first year, their average completion time does not reduce.

To sum it up, performance standards substantially increase dropout after the first year. About half of these additional dropouts would have graduated without the policy. The average enrollment duration declines, reflecting earlier exits of students who would otherwise have remained enrolled, and in some cases graduated, rather than faster completion among stayers.

Table 4: The effect of performance standards on outcomes in the initial program

	Dropout after year two	Dropout after year three or later	Complete initial program	Years enrolled in initial program	Completion time (conditional on graduation)
	(1)	(2)	(3)	(4)	(5)
ATT	-0.022***	-0.018***	-0.036***	-0.237***	-0.042
	(0.003)	(0.003)	(0.006)	(0.039)	(0.032)
ATT_{-1}	0.001	0.005	-0.010	-0.028	-0.006
	(0.004)	(0.003)	(0.006)	(0.034)	(0.031)
ATT_{-2}	0.002	0.006	-0.010	-0.047	-0.045
	(0.005)	(0.005)	(0.011)	(0.055)	(0.046)
ATT_{-3}	0.006	0.016**	-0.021*	0.002	0.009
	(0.006)	(0.007)	(0.012)	(0.065)	(0.051)
ATT_{-4}	0.003	0.011*	-0.012	-0.004	-0.018
	(0.006)	(0.006)	(0.011)	(0.068)	(0.070)
Mean	0.071	0.061	0.639	3.292	4.255
SD	(0.041)	(0.047)	(0.125)	(0.626)	(0.890)
N	7,579	7,579	7,579	7,579	7,579

Notes: Each column presents event-study estimates of the effect of implementing performance standards on a different outcome. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the dynamic effects for the first three periods after treatment. See Figure A5 for all event-study estimates. The next four rows present placebo estimates based on pre-treatment period differences. The last two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

5.3 Long-term effects

Higher education. Table 5 column 1 shows that implementing performance standards reduces graduation from any field at a research university by 1.4 percentage points.[27] I find a similar but less precisely estimated effect when I also include degrees from University of Applied Sciences in column 2. This reduction is lower than the reduction in the graduation rate from the initial program (3.6 percentage points), which implies that some of the students that dropped out from the initial program manage to graduate elsewhere. However, the goal of performance standards was to redirect students towards programs where they were *more* likely to graduate. The results suggest that performance standards do not achieve this goal.

Performance standards were also intended to reduce the time that students spend in higher education by redirecting students towards more suitable programs at an early stage. However, column 3 shows that the average time spent in higher education is unaffected. The estimate is precise and rules out enrollment duration reductions by 0.14 years (2 percent) with 95 percent confidence. This shows that although the average enrollment duration in the initial program is reduced, dismissed students often re-enroll elsewhere and spend as much time in higher education as they would have done otherwise. I also find no effects on the 90th percentile of the enrollment duration distribution (column 4), suggesting no visible impact even among the slowest students.

Labor market outcomes. I next study effects on post-study earnings. This dimension is relevant because the higher education outcomes above may not fully reflect students' human capital accumulation. For example, even when the effects on higher education attainment and time in higher education are small, students' labor market outcomes may have improved if they shifted towards more suitable careers or if their performance improved in a way that is not reflected in graduation rates and enrollment durations.

[27] In the last pre-treatment period there is a small and marginally significant deviation from zero. This pattern also appears in column 2 because the outcome is highly correlated with the outcome in column 1. Such minor deviations are expected under random variation. Table A5 shows that the estimate is stable when programs are matched on pre-trends. This suggests that this small deviation does not drive the post-treatment period ATT estimate.

Table 5: The effect of performance standards on long-term outcomes

	Obtain research university degree	Obtain higher education degree	Years enrolled in higher education	Years enrolled in higher education (90th quantile)	Annual earnings 12 years post enrollment (€)
	(1)	(2)	(3)	(4)	(5)
ATT	-0.014**	-0.013*	-0.060	-0.015	191
	(0.006)	(0.007)	(0.040)	(0.057)	(372)
ATT_{-1}	-0.008*	-0.007*	0.010	0.052	-423
	(0.005)	(0.004)	(0.034)	(0.087)	(467)
ATT_{-2}	0.003	0.004	0.005	0.025	58
	(0.006)	(0.004)	(0.039)	(0.088)	(502)
ATT_{-3}	-0.004	-0.003	0.006	0.042	-228
	(0.009)	(0.007)	(0.051)	(0.097)	(639)
ATT_{-4}	-0.001	-0.004	0.013	0.119	-286
	(0.008)	(0.007)	(0.057)	(0.106)	(617)
Mean	0.805	0.885	6.245	8.981	63,577
SD	(0.081)	(0.063)	(0.802)	(1.075)	(14,421)
N	7,579	7,579	7,579	7,579	6,531

Notes: Each column presents event-study estimates of the effect of implementing performance standards on a different outcome. With the exception of column 4 and 5, the program-level outcomes are averaged over all students who initially enrolled in the respective program. The outcome in column 4 is the 90th quantile of the enrollment duration distribution, which is also taken over all initially enrolled students. Income in column 5 is taken over all initially enrolled students with observed and positive income. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the dynamic effects for the first three periods after treatment. See Figure A6 for all event-study estimates. The next four rows present placebo estimates based on pre-treatment period differences. The last two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. P-values are adjusted for multiple hypothesis testing using simultaneous confidence bands. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

However, I find no evidence for labor market improvements. Because earnings twelve years after enrollment is only available for the earlier cohorts, I focus on students enrolled before 2012 (see Table 1). I first test whether implementing performance standards affects whether students' earnings twelve years after the initial enrollment is observed and positive in Table A4.[28] The estimates are insignificant, suggesting that selection into observed earnings due to performance standards is unlikely to be an issue. I then estimate the effect on earnings twelve years post first enrollment. Table 5 column 5 shows a precisely estimated effect around zero, ruling out earnings increases of more than 1.4 percent with 95 percent certainty.

Discussion. Table A5 shows that the main results are similar when using a Synthetic Difference-in-Differences approach, which is consistent with only minor deviations in pre-trends even without matching. Table A6 shows that the results also do not change when excluding international students, whose visa depends on continued enrollment and for whom dismissal is therefore more consequential.

Overall, the results suggest that performance standards do not meaningfully improve students' average education or labor market outcomes. The next section examines whether these average effects mask heterogeneous effects.

5.4 Heterogeneity

To limit the number of hypothesis tests and keep the results concise, I focus on two outcomes: (i) the dropout rate after the first year, to explore which groups are most impacted by performance standards in the short run, and (ii) the average enrollment duration per graduate. The latter is defined as the average number of years that all initially enrolled students are enrolled in higher education divided by their average graduation rate in higher education. The advantage of this index is that it summarizes long-run effects on two key outcomes into

[28] Table A4 also reports effects for two other outcomes: first, I show that the impact of the performance standard on the dropout rate is similarly large for these cohorts, suggesting that the treatment is comparable for this period. Second, I find no effects on migration, suggesting that differential migration due to performance standards is no issue.

one measure, and it can be interpreted as a crude cost benefit ratio. If performance standards reduce enrollment durations (costs) or increases graduation rates (benefits), then this ratio should decrease. Figure 4 reports estimates for these outcomes for several subgroups.

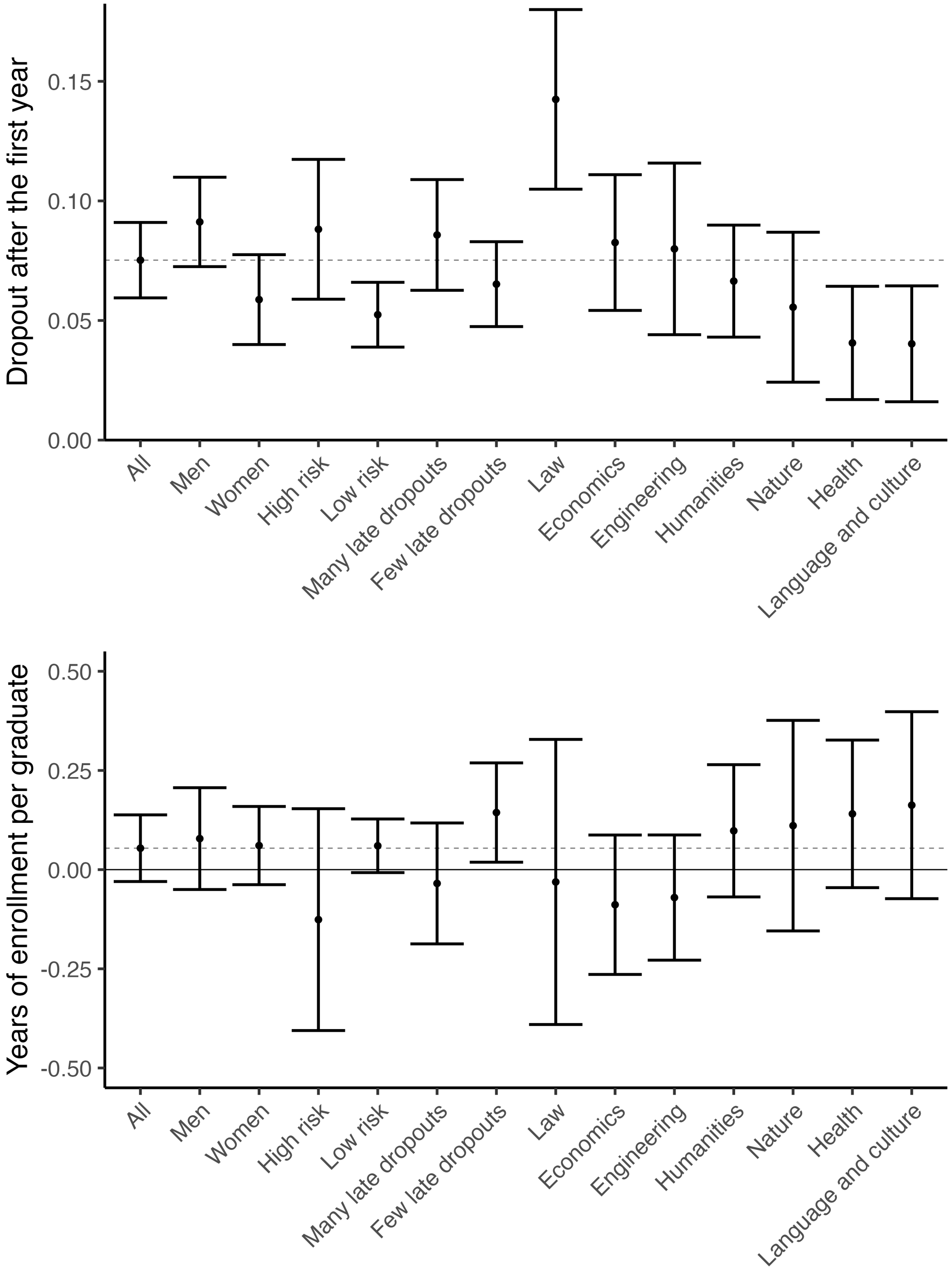
I start by exploring heterogeneity by gender because males perform worse in higher education and drop out more often. A plausible explanation is that males are less patient, leading them to underinvest in their studies (Castillo et al. (2011)). By imposing an incentive to perform early in the program, performance standards could therefore be more beneficial for men. Consistent with lower performance among men, Figure 4 shows that men are more likely to be dismissed. However, in the long run, there is no improvement in the average enrollment duration per graduate for men or women.

Second, I test whether performance standards differentially affect high-risk students, where high-risk students are defined as the 25 percent students in each program with the lowest predicted on-time graduation probability based on their pre-enrollment characteristics (as in Section 5.1). As expected, these students are more likely to be dismissed. However, also for this group, I find no long-run effects.

Next, I examine heterogeneity across two program characteristics. First, I consider programs with a high share of late dropouts, defined as programs with an above median share of students dropping out after the second year or later in the pre-treatment periods. Since performance standards were intended to redirect unfit students into more suitable programs at an early stage, these programs offer the greatest scope for improvement. However, Figure 4 shows that while dismissals are indeed somewhat more frequent in these programs, there are no effects on long-run outcomes for this group neither. For programs with relatively few late dropouts, performance standards even slightly increase the average enrollment duration per graduate. I next consider heterogeneity by field of study. Although short-run effects vary somewhat across fields, the long-run impacts are also consistently small.

Finally, I assess heterogeneity across implementation cohorts in Figure A7. I find that the dismissal rate is somewhat higher among the early adopters, and this also results in

Figure 4: Heterogeneous effects of performance standards



Note: This figure presents estimates of the effect of implementing performance standards on first-year dropout and the average enrollment duration in higher education per graduate for several subgroups. The estimates by gender and risk status are obtained by splitting the individual data by gender or risk status, collapse the data to the program level for each group separately, and apply difference-in-differences to each dataset separately. The estimates by the field of study are obtained as follows: (i) select all treated and untreated observations for programs from a specific field and select all untreated observations from all other programs, (ii) apply difference-in-differences to this subsample to obtain the field specific ATT, and (iii) repeat this process for all other fields. The same steps are used for estimating heterogeneity by the share of late dropouts in a program. The dashed line is centered around the ATT for the full sample. The error bars present 95% confidence intervals.

longer-run impacts. However, these effects correspond only to a small number of programs (15 percent), are mostly negative, and are quite imprecisely estimated. For the larger share of programs that implemented performance standards after 2003, the average enrollment duration per graduate is not affected.

To conclude, while the short-run effects vary somewhat between groups, the long-run effects are consistently small. Performance standards do not measurably improve the average enrollment duration per graduate for males or females, low- or high-risk students, programs with a high or low share of late dropouts, programs in specific fields, and - with the exception of a small share of early adopters - the effects are consistently small across implementation cohorts.

6 The non-monetary costs of performance standards

6.1 Measuring (dis)utility with survey evidence

Performance-based dismissal policies restrict the failing students' choices by taking away the option to continue in their program. In the absence of spillovers or behavioral biases, reducing choice sets is unlikely to increase welfare. The intervention is not free of monetary costs either, because universities face administrative costs and appeal cases.

These costs can be justified when performance standards reduce costs imposed on others. For example, the goal was to reduce the time students spend in education, thereby reducing education subsidies and forgone earnings. However, the results indicate that performance standards do not improve students' time spent in education or labor market outcomes.

Another justification could be an increase in positive internalities (Chetty (2015)). For example, adolescents are known to disproportionally discount the future benefits of education (Bleemer and Zafar (2018)), inducing them to invest suboptimally in their studies. If that is the case for first-year bachelor students too, then the performance standard can be a welcome commitment device that gives students a more immediate incentive to increase their effort level (Clark et al. (2020)).

To test whether the performance standard is perceived as a distortionary intervention or a welcome commitment device, I surveyed 333 first-year students in two bachelor programs from a large university in the Netherlands. The survey is conducted in the first semester, when the students finished their first exams but have not received grades yet. They are asked five versions of the same hypothetical choice question where they trade off monetary gifts and the immediate removal of the performance standard.[29] I use a staircase method that enables me to narrow down the interval around the amount where each student is exactly indifferent between the gift and the removal of the performance standard.[30] Importantly, for students with a sufficiently low willingness to forgo gifts, I also present the option to *pay* a small amount of money to keep the performance standard in place. This option corresponds to students who appreciate the presence of a performance standard.

An advantage of the hypothetical choice approach is that it enables flexible estimation of the full distribution of preferences (Wiswall and Zafar, 2018). This allows me to detect both individuals who appreciate and who dislike the performance standard. Even if administrative data included measures of student utility or satisfaction—which it does not—it would remain difficult to capture such preference heterogeneity.

Most students experience disutility from performance standards. Figure 5(a) shows the distribution of students' willingness to forgo money for their immediate removal. A minority (18 percent) would pay to retain the standard, suggesting that for some it serves as a commitment device. Yet a much larger share is willing to forgo money to remove it. For example, 24 percent would give up between 500 and 1700 euros, and 26 percent indicate the maximum of 1700 euros. This far exceeds the average monthly disposable income of students, estimated at 943 euros (Groen and Houtsma (2021)).

While the willingness to forgo gifts increases with the subjective risk of dismissal, even

[29] The exact phrasing of the question and other details of the survey are discussed in Appendix D.

[30] Similar staircase methods are commonly used to calculate individuals' certainty equivalents or individual time discounting rates (Falk et al. (2018)).

students who perceive little risk are often willing to forgo gifts. To measure this risk, I elicit students' expectations about the number of course credits they will have by the end of the year and use these to construct the implied probability of dismissal.[31] Figure 5(b) shows boxplots of the willingness to forgo gifts against these subjective probabilities. The median willingness to forgo gifts increases steeply with dismissal risk, suggesting that students most likely to fail the standard experience the strongest disutility. Yet the median willingness remains positive in all four probability bins, including among students who are certain they will have sufficiently many credits. This indicates that the disutility from performance standards extends beyond those at risk of dismissal.[32]

Importantly, willingness to forgo money is hypothetical and elicited before students fully experience the long-run consequences of the policy. Students who dislike the performance standard in the first year may revise their views if they later believe it improved their performance or led to a better matched program. The results therefore warrant caution. Nevertheless, the magnitude and pervasiveness of the reported disutility are notable. This suggests that performance-based dismissal policies generate costs that are invisible in administrative outcomes but likely relevant for student welfare.

6.2 Mental health

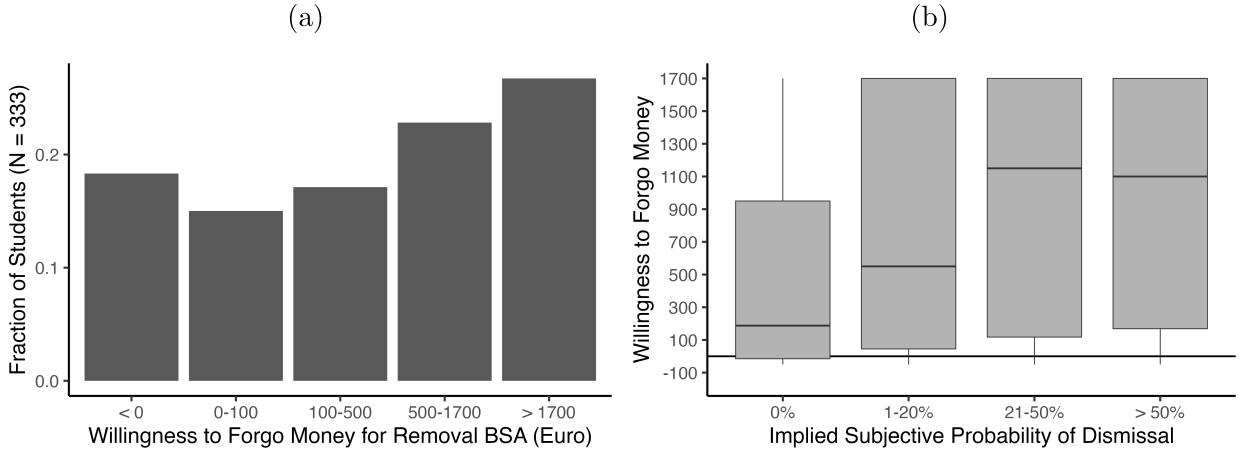
A specific way through which performance standards can reduce students' utility is if it unintentionally impacts their mental well-being. Student mental health in the Netherlands has shown a declining trend, and the binding study advice is repeatedly mentioned as a potential contributor (Dopmeijer et al. (2022)). Despite anecdotal evidence linking mental health concerns to performance standards, there exists no causal evidence on its effect.

To shed light on this, I study the effect of performance standards on antidepressant

[31] Specifically, instead of asking about their subjective likelihood of dismissal directly, I follow Manski (2004), and ask students to allocate 100 points over each possible number of credits they can obtain. I then define a student's subjective probability of being dismissed by the implied probability that a student has fewer credits than necessary to pass the performance standard.

[32] A plausible explanation is that these students expect to meet the credit requirement if they apply sufficient effort, but would prefer to exert less effort than the performance standard demands.

Figure 5: Willingness to forgo gifts for removal of the performance standard



Note: figure (a) presents a bar chart of students' willingness to forgo gifts in return for the immediate removal of the performance standard. The willingness to forgo money is based on survey responses to five hypothetical choice questions, discussed in further detail in Appendix D. Figure (b) shows boxplots of the willingness to forgo money for four groups of students, sorted based on their implied subjective probability of dismissal. The boxes mark the interquartile range, the line inside marks the median, and whiskers extend to 1.5 times the interquartile range. The share of students in the bins is 37% (bin 0), 24% (bin 1-20), 19% (bin 21-50), and 20% (bin > 50).

prescriptions. Because prescription data are available only from 2007 and uptake in the first year of enrollment is low (2.6 percent), I cannot adequately study effects during the first year. I instead examine antidepressant uptake between 5 and 8 years after the first enrollment.[33] Such effects could reflect dissatisfaction among dismissed students with their new career paths. I find no evidence that performance standards raise antidepressant use. Table A7 reports a small and insignificant estimate and rules out an increase larger than 0.7 percentage points (11 percent) with 95 percent confidence.

While this result is a step towards a better understanding of mental health effects, they do not rule out negative effects on less extreme mental health outcomes. Performance standards may affect students' mental health in other ways that do not directly manifest in antidepressant usage.

[33] Restricting to uptake after 5 years allows me to include all programs that introduced a performance standard after 2002, while limiting the window to 8 years ensures complete coverage of prescriptions for all adoption cohorts.

7 Conclusion and discussion

This paper evaluates the introduction of strict performance standards in Dutch bachelor programs between 1994 and 2014. The performance standards were intended to improve student outcomes by incentivizing effort in the first year and redirecting students unlikely to succeed toward more suitable career paths. My results suggest that these objectives are largely not achieved.

In the short-run, performance standards result in the dismissal of many students from their preferred programs. In the long run, they slightly reduce degree attainment without shortening time spent in higher education, and they do not improve labor market outcomes. Survey evidence further indicates that performance standards cause disutility.

Taken together, the findings cast doubt on the effectiveness of performance induced dismissal as a tool to improve student outcomes. The Dutch higher education system offers relatively generous outside options for dismissed students, who can re-enroll in many alternative programs at low costs. Even in this environment, the policy fails to generate benefits. This raises concerns about similar policies in less forgiving contexts. Where tuition is higher, admission constraints are tighter, or nearby alternatives are scarce, dismissal may push students out of higher education entirely, with potentially sizeable earnings losses as a result (Ost et al. (2018)). In such settings, any gains among continuing students would need to be large to compensate for losses among those who are dismissed.

Lastly, the estimates capture average effects across all initially enrolled students and focus on long-run outcomes such as degree attainment, enrollment duration, and earnings. This approach provides insights on whether performance standards ultimately improve aggregate educational and labor market outcomes that policymakers care about. However, it does not allow for a detailed examination of the mechanisms underlying the absence of long run effects. Because grades and credits are not observed, I cannot assess whether initial performance gains faded over time or distinguish effects between continuing and dismissed students. Some students who met the thresholds may have benefited in the long run, but such gains could be offset by losses among dismissed students. These mechanisms remain an important direction for future research.

References

Aizer, Anna, Nancy Early, Shari Eli, Guido Imbens, Keyoung Lee, Adriana Lleras-Muney, and Alexander Strand. 2024. “The Lifetime Impacts of the New Deal's Youth Employment Program.” *The Quarterly Journal of Economics* 139 (4): 2579–2635.

Al-Ubaydli, Omar, John A. List, Danielle LoRe, and Dana Suskind. 2017. “Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature.” *Journal of Economic Perspectives* 31 (4): 125–144.

Albert, Aaron, and Nathan Wozny. 2024. “The Impact of Academic Probation: Do Intensive Interventions Help?” *Journal of Human Resources* 59 (3): 852–878.

Almond, Douglas, Janet Currie, and Valentina Duque. 2018. “Childhood Circumstances and Adult Outcomes: Act II.” *Journal of Economic Literature* 56 (4): 1360–1446.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. “Incentives and Services for College Achievement: Evidence from a Randomized Trial.” *American Economic Journal: Applied Economics* 1 (1): 136–163.

Angrist, Joshua, and Victor Lavy. 2009. “The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial.” *American Economic Review* 99 (4): 1384–1414.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. 2021. “Synthetic Difference-in-Differences.” *American Economic Review* 111 (12): 4088–4118.

Arnold, Ivo J.M. 2015. “The Effectiveness of Academic Dismissal Policies in Dutch University Education: An Empirical Investigation.” *Studies in Higher Education* 40 (6): 1068–1084.

Avdeev, Stanislav, Nadine Ketel, Hessel Oosterbeek, and Bas van der Klaauw. 2024. “Spillovers in Fields of Study: Siblings, Cousins, and Neighbors.” *Journal of Public Economics* 238 105193.

Bleemer, Zachary, and Basit Zafar. 2018. “Intended College Attendance: Evidence from an Experiment on College Returns and Costs.” *Journal of Public Economics* 157 184–211.

Bursztyn, Leonardo, and Robert Jensen. 2015. “How Does Peer Pressure Affect Educational Investments?” *The Quarterly Journal of Economics* 130 (3): 1329–1367.

Callaway, Brantly, and Pedro H.C. Sant'Anna. 2021. “Difference-in-Differences with Multiple Time Periods.” *Journal of Econometrics* 225 (2): 200–230.

Canaan, Serena, Stefanie Fischer, Pierre Mouganie, and Geoffrey C. Schnorr. 2023. “Keep Me In, Coach: The short- and long-term effects of targeted academic coaching.” CLEF Working Paper Series 60, Canadian Labour Economics Forum (CLEF), University of Waterloo, `https://ideas.repec.org/p/zbw/clefwp/60.html`.

Casey, Marcus D., Jeffrey Cline, Ben Ost, and Javaeria A. Qureshi. 2018. “Academic Probation, Student Performance, and Strategic Course-Taking.” *Economic Inquiry* 56 (3): 1646–1677.

Castillo, Marco, Paul J. Ferraro, Jeffrey L. Jordan, and Ragan Petrie. 2011. “The Today and Tomorrow of Kids: Time Preferences and Educational Outcomes of Children.” *Journal of Public Economics* 95 (11): 1377–1385.

Caves, Katherine, and Simone Balestra. 2018. “The Impact of High School Exit Exams on Graduation Rates and Achievement.” *The Journal of Educational Research* 111 (2): 186–200.

Chetty, Raj. 2015. “Behavioral Economics and Public Policy: A Pragmatic Perspective.” *American Economic Review* 105 (5): 1–33.

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush. 2020. “Using Goals to Motivate College Students: Theory and Evidence From Field Experiments.” *The Review of Economics and Statistics* 102 (4): 648–663.

Clark, Damon, and Paco Martorell. 2014. “The Signaling Value of a High School Diploma.” *Journal of Political Economy* 122 (2): 282–318.

de Chaisemartin, Clément, and Xavier D’Haultfœuille. 2020. “Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects.” *American Economic Review* 110 (9): 2964–2996.

Dopmeijer, JM, J Nuijen, MJM Busch, NI Tak, and N van Hasselt. 2022. “Monitor Mentale gezondheid en Middelengebruik Studenten hoger onderwijs. Deelrapport II. Middelengebruik van studenten in het hoger onderwijs.” 10.21945/RIVM-2022-0101.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. “Global Evidence on Economic Preferences.” *The Quarterly Journal of Economics* 133 (4): 1645–1692.

Fidjeland, Andreas. 2023. “Using High-Stakes Grades to Incentivize Learning.” *Economics of Education Review* 94 102377.

Figlio, David, and Susanna Loeb. 2011. “Chapter 8 - School Accountability.” In *Handbook of the Economics of Education*, edited by Hanushek, Eric A., Stephen Machin, and Ludger Woessmann Volume 3. 383–421.

Fletcher, Jason M., and Mansur Tokmouline. 2018. “The Effects of Academic Probation on College Success: Regression Discontinuity Evidence from Four Texas Universities.” IZA Discussion Paper No. 11232.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. 2021. “Visualization, Identification, and Estimation in the Linear Panel Event-Study Design.” National Bureau of Economic Research Working Paper Series No. 29170.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29 (5): 1189–1232.

Groen, Annette, and Nanne Houtsma. 2021. “Nibud Studentenonderzoek 2021.” `https://www.nibud.nl/onderzoeksrapporten/nibud-studentenonderzoek-2021/`, Accessed: 31-10-2025.

Hvidman, Ulrik, and Hans Henrik Sievertsen. 2021. “High-Stakes Grades and Student Behavior.” *Journal of Human Resources* 56 (3): 821–849.

Lavecchia, A.M., H. Liu, and P. Oreopoulos. 2016. “Behavioral Economics of Education.” In *Handbook of the Economics of Education*, Volume 5. 1–74.

Leuven, Edwin, Hessel Oosterbeek, Joep Sonnemans, and Bas van der Klaauw. 2011. “Incentives versus Sorting in Tournaments: Evidence from a Field Experiment.” *Journal of Labor Economics* 29 (3): 637–658.

Leuven, Edwin, Hessel Oosterbeek, and Bas Van der Klaauw. 2010. "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment." *Journal of the European Economic Association* 8 (6): 1243–1265.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Economic Journal: Economic Policy* 8 (4): 183–219.

Lindo, Jason M., Nicholas J. Sanders, and Philip Oreopoulos. 2010. "Ability, Gender, and Performance Standards: Evidence from Academic Probation." *American Economic Journal: Applied Economics* 2 (2): 95–117.

Manski, Charles F. 2004. "Measuring Expectations." *Econometrica* 72 (5): 1329–1376.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister et al. 2024. "Nonstandard Errors." *The Journal of Finance* 79 (3): 2339–2390.

Ministerie van Onderwijs, Cultuur en Wetenschap. 2010. "Met beide benen op de grond: onderzoek naar uitvoeringspraktijk bindend studieadvies in hoger onderwijs." rapport, Den Haag.

OECD. 2025. "Education at a Glance 2025: OECD Indicators." 10.1787/1c0d9c79-en.

Ost, Ben, Weixiang Pan, and Douglas Webber. 2018. "The Returns to College Persistence for Marginal Students: Regression Discontinuity Evidence from University Dismissal Policies." *Journal of Labor Economics* 36 (3): 779–805.

Ou, Dongshu. 2010. "To Leave or Not to Leave? A Regression Discontinuity Analysis of the Impact of Failing the High School Exit Exam." *Economics of Education Review* 29 (2): 171–186.

Papay, John P., Richard J. Murnane, and John B. Willett. 2016. "The Impact of Test Score Labels on Human-Capital Investment Decisions." *Journal of Human Resources* 51 (2): 357–388.

Porreca, Zachary. 2022. "Synthetic Difference-in-Differences Estimation with Staggered Treatment Timing." *Economics Letters* 220 110874.

Rodríguez-Planas, Núria. 2012. "Longer-Term Impacts of Mentoring, Educational Services, and Learning Incentives: Evidence from a Randomized Trial in the United States." *American Economic Journal: Applied Economics* 4 (4): 121–139.

Roth, Jonathan. 2022. "Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends." *American Economic Review: Insights* 4 (3): 305–322.

Sneyers, Eline, and Kristof De Witte. 2017. "The Effect of an Academic Dismissal Policy on Dropout, Graduation Rates and Student Satisfaction. Evidence from the Netherlands." *Studies in Higher Education* 42 (2): 354–389.

ter Meulen, Simon. 2023. "Long-Term Effects of Grade Retention." CESifo Working Paper No. 10212, Center for Economic Studies and ifo Institute (CESifo), Munich.

Universiteiten van Nederland. 2023. "Bindend studieadvies is goed voor student en studentsucces." `https://www.universiteitenvannederland.nl/onderwerpen/onderwijs/bindend-studieadvies-is-goed-voor-student-en-studentsucces`, Accessed January 19, 2026.

Vooren, Melvin, Ilja Cornelisz, Martijn Meeter, and Chris Van Klaveren. 2024. "Abolishing Policy-Induced Dropout in Higher Education Due to Covid-19. Exploring Trends in Future Academic Performance." SSRN Working Paper No. 4958398, SSRN.

Wiswall, Matthew, and Basit Zafar. 2018. "Preference for the Workplace, Investment in Human Capital, and Gender." *The Quarterly Journal of Economics* 133 (1): 457–507.

Wright, Nicholas A. 2020. "Perform Better, or Else: Academic Probation, Public Praise, and Students Decision-Making." *Labour Economics* 62 101773.

Appendix A: Supplementary Tables and Figures

Table A1: The adoption of performance standards by university

University	1995	1996	1999	2001	2002	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
1											3		1	13	2	
2			1				14									
3	5	33														
4												43				
5					4	2		1					1	1	25	13
6											13	1				
7											11					
8					2	3	10	17	9	1						
9						2			3		1	34	1			
10					8	6										
11			4	3		1	1				1	1				
12													37			
13																19

Notes: This table reports for each of the thirteen universities how many programs adopted a performance standard in each year.

Table A2: The effect of program characteristics on the adoption timing of performance standards

	Year of adoption performance standard
log(programSize)	-0.256
	(0.325)
Graduation rate	1.928
	(3.328)
Average enrollment duration	-0.865
	(0.690)
Field = agriculture	6.257*
	(3.221)
Field = Life sciences	0.981
	(3.122)
Field = Engineering	2.781
	(3.196)
Field = Health	1.576
	(3.147)
Field = Economics	-2.249
	(3.173)
Field = Law	-2.430
	(3.270)
Field = Humanities	0.154
	(3.047)
Field = Language and culture	-0.771
	(3.117)
N	351

Notes: This table presents results from a regression of the year of adoption of a performance standard on the log of the average program size, the average secondary school GPA, the average graduation rate, and indicators for the field of study. These variables are averaged over all enrolled students in each program from the core sample. Since grades are available only since 2007, grade information is averaged only over the later years. The left-out indicator for field of study corresponds to Education Sciences. The sample includes all but one program from the main analysis. One program is missing because it implemented the performance standard in 1996 and ceased to exist before 2007, when the first grade information becomes available. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Table A3: Effect of performance standards on average secondary school grades of new students using synthetic difference-in-differences

	Secondary school GPA
ATT	0.001
	(0.009)
N	2,792

Notes: This table presents a synthetic difference-in-difference (SDID) estimate of the effect of performance standards on the average secondary school GPA of new students. Since grades are only available since 2007 and SDID requires at least two pre-treatment periods, the treatment group only includes programs that implemented a performance standard after 2008. The estimates are obtained by applying SDID to each adoption cohort separately, and then aggregating their results (Arkhangelsky et al. (2021)). The estimates are weighted by the average program size in the pre-treatment period. The standard errors follow from Porreca (2022). (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Table A4: Testing for selectivity into observed earnings using the income-analysis sample

	Dropout after first year	Missing income	Positive income	Migration
ATT	0.083***	0.011	0.002	-0.001
	(0.009)	(0.008)	(0.002)	(0.005)
ATT_{-1}	0.005	0.007	0.002	0.005
	(0.007)	(0.005)	(0.003)	(0.004)
ATT_{-2}	0.004	-0.005	0.004	-0.002
	(0.009)	(0.006)	(0.003)	(0.006)
ATT_{-3}	0.001	-0.007	0.002	-0.004
	(0.010)	(0.012)	(0.003)	(0.009)
ATT_{-4}	-0.001	-0.007	0.002	-0.004
	(0.009)	(0.013)	(0.003)	(0.009)
Mean	0.234	0.153	0.958	0.127
SD	(0.090)	(0.155)	(0.028)	(0.110)
N	6,531	6,531	6,531	6,531

Notes: Each column presents event-study estimates of the effect of implementing performance standards on a different outcome. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the ATTs for three post-treatment periods Callaway and Sant'Anna (2021). The next four rows present placebo estimates based on pre-treatment period differences. The last two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Table A5: Main results, using Synthetic difference-in-difference

	New enrollments	Dropout after first year	Complete initial program	Years enrolled in initial program
ATT	-2.90*	0.070***	-0.027***	-0.230***
	(1.74)	(0.011)	(0.006)	(0.046)
N	7,579	7,579	7,579	7,579

	Obtain research university degree	Obtain higher education degree	Years enrolled in higher education	Annual income 12 years post enrollment
ATT	-0.015***	-0.013**	-0.062	328
	(0.005)	(0.006)	(0.043)	(779)
N	7,579	7,579	7,579	6,531

Notes: This table presents synthetic difference-in-difference estimates of the effect of performance standards on various outcomes. The estimates are obtained by applying synthetic difference-in-differences to each adoption cohort separately, and then aggregating their results (Arkhangelsky et al. (2021)). All results are based on the same samples as in the main results and the estimates are weighted by the average program size in the pre-treatment period. The standard errors follow from Porreca (2022). (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Table A6: Main results, excluding international students

	New enrollments	Dropout after first year	Complete initial program	Years enrolled in initial program
ATT	-3.08	0.076***	-0.038***	-0.240***
	(3.39)	(0.007)	(0.009)	(0.057)
N	7,549	7,549	7,549	7,549

	Obtain research university degree	Obtain higher education degree	Years enrolled in higher education	Annual income 12 years post enrollment
ATT	-0.014**	-0.013*	-0.060	88
	(0.007)	(0.007)	(0.043)	(469)
N	7,549	7,549	7,549	6,503

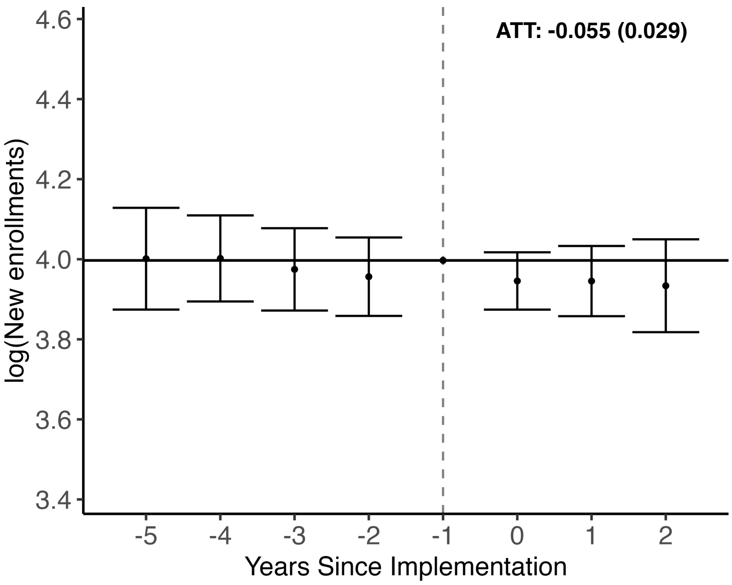
Notes: This table replicates the analysis from section 5. The only difference is that the sample here excludes all international students, who represent 11 percent of the core analysis sample. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Table A7: The effect of performance standards on antidepressant prescriptions

	Antidepressant prescription within five to eight years of enrollment
ATT	-0.001
	(0.004)
ATT_{-1}	0.004
	(0.005)
ATT_{-2}	0.004
	(0.005)
ATT_{-3}	0.003
	(0.006)
ATT_{-4}	0.007
	(0.006)
Mean	0.062
SD	(0.039)
N	4,493

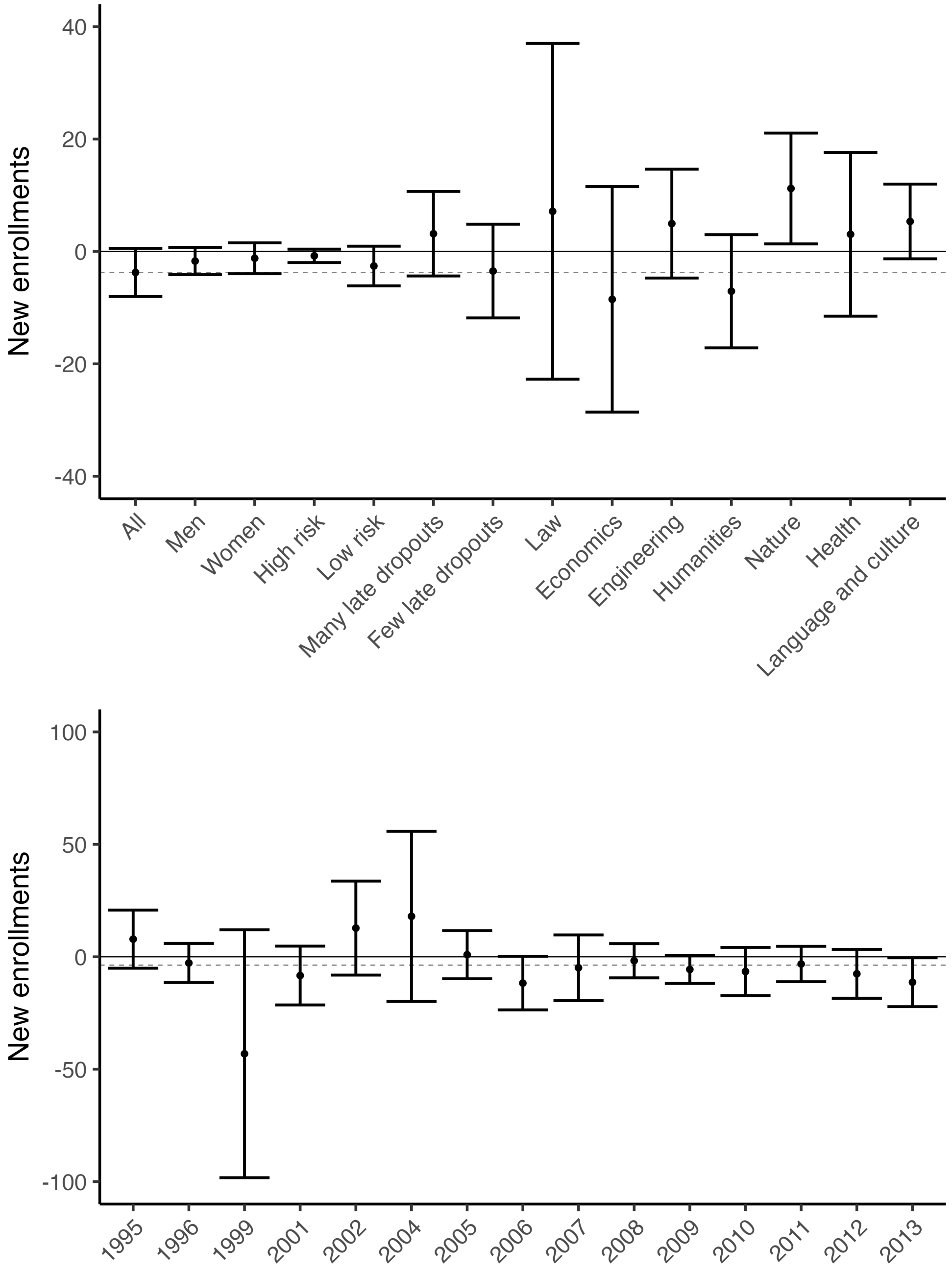
Notes: This table presents event-study estimates of the effect of implementing performance standards on the antidepressant prescriptions within five years of the initial enrollment. Row 1 (ATT) presents the Average Treatment effect on the Treated, which is a weighted average of the ATTs for three post-treatment periods Callaway and Sant'Anna (2021). The next two rows present placebo estimates based on pre-treatment period differences. The last two rows report the mean and standard deviation of each outcome in the last pre-treatment period. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Figure A1: The effect of performance standards on new enrollments in logs



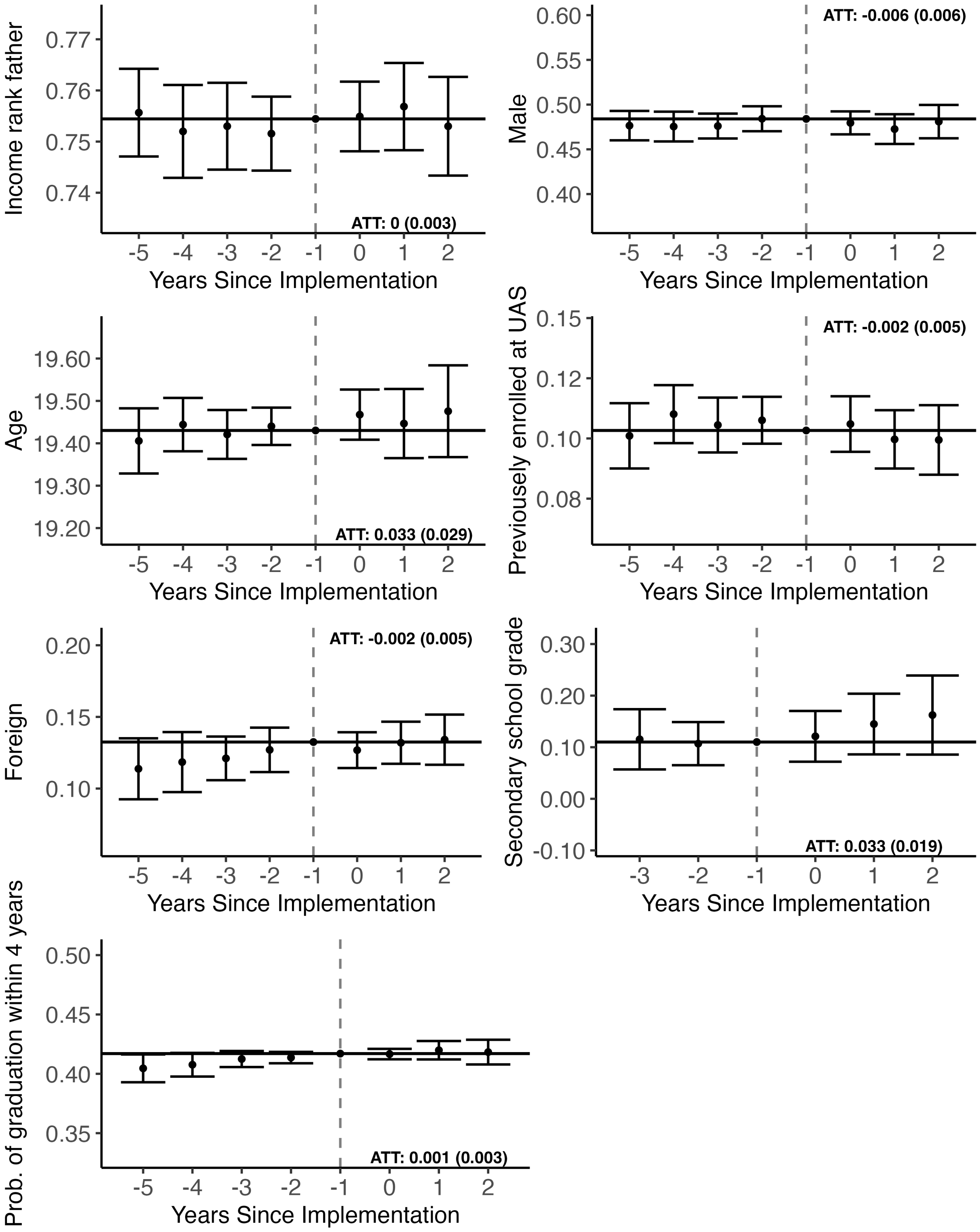
Note: This figure presents event-study estimates of the effect of implementing performance standards on new enrollments. The sample comprises of all 7,579 observations between 1994 and 2014 of all 351 programs. The estimates are obtained using Callaway and Sant'Anna (2021). The ATT is based on the weighted average of the dynamic effects (β_l) for the first three years after implementation ($l \in \{0, 1, 2\}$). The graph is centered around the average program size in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing.

Figure A2: Effects of implementing performance standards on program size across subgroups and implementation cohorts



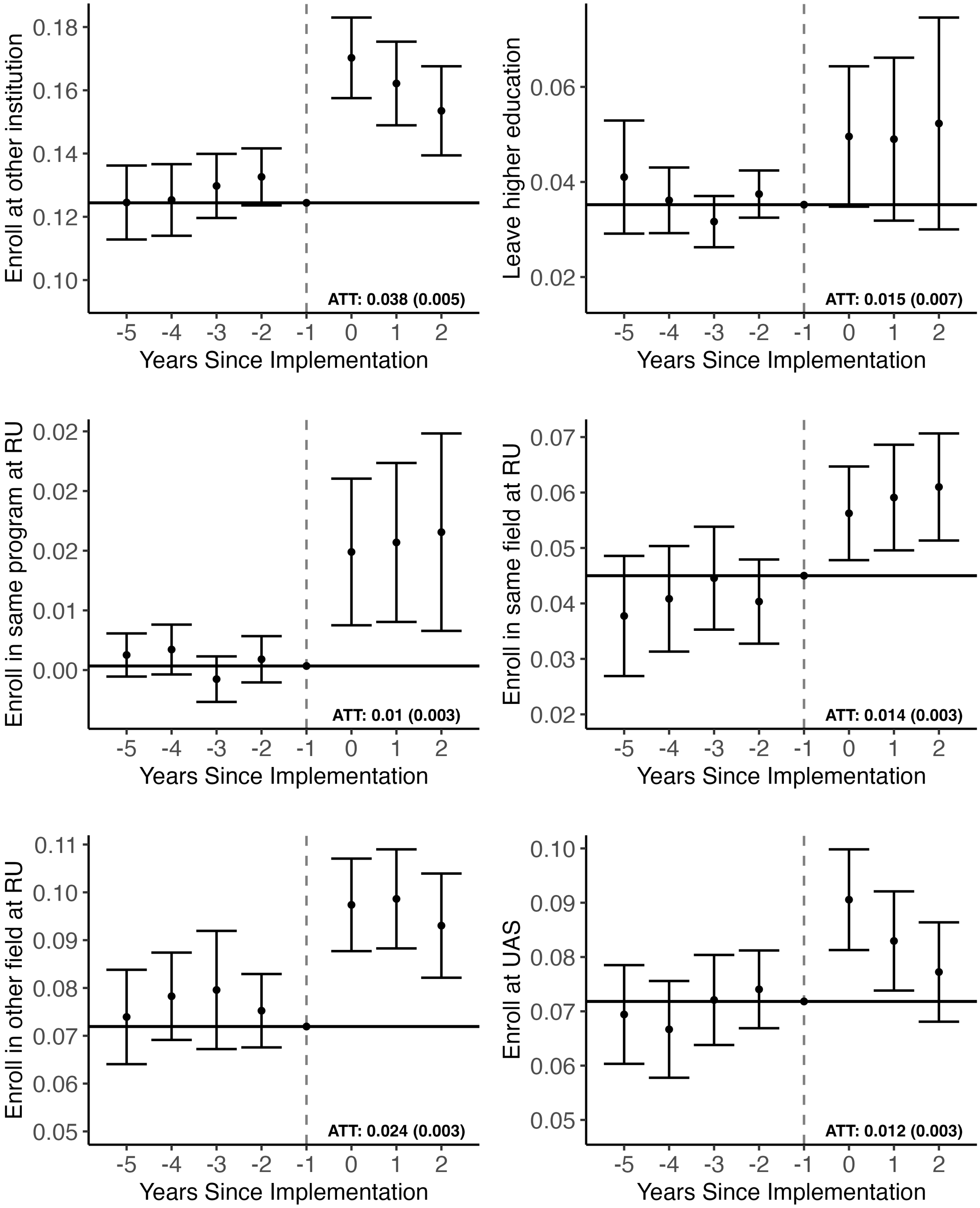
Note: This figure presents estimates of the effect of implementing performance standards on new enrollments across several subgroups (upper panel) and across all adoption cohorts (lower panel). The estimates by gender and risk status are obtained by splitting the individual data by gender or risk status, collapse the data to the program level for each group separately, and apply difference-in-differences to each dataset separately. The estimates by the field of study are obtained as follows: (i) select all treated and untreated observations for programs from a specific field and select all untreated observations from all other programs, (ii) apply difference-in-differences to this subsample to obtain the field specific ATT, and (iii) repeat this process for all other fields. The same steps are used for estimating heterogeneity by the share of late dropouts in a program. The lower-panel estimates are constructed using a group-level decomposition of the full sample ATT (Callaway and Sant'Anna (2021)). The dashed lines are centered around the ATT for the full sample. The error bars present 95% confidence intervals.

Figure A3: Event study estimates from Table 2



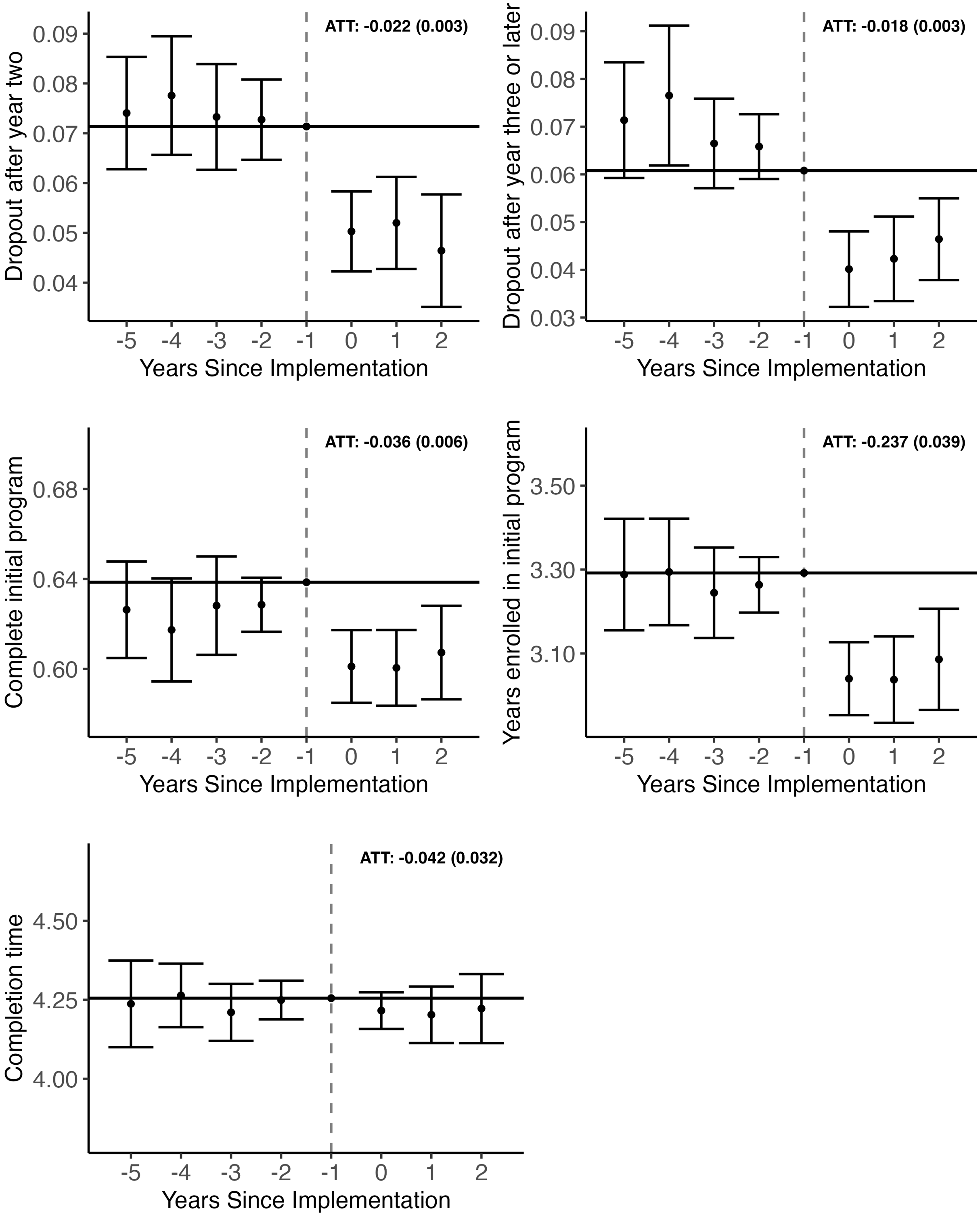
Note: This figure presents the event-study estimates corresponding to the results in Table 2. The graphs are centered around the average of the outcome in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing.

Figure A4: Event study estimates from Table 3



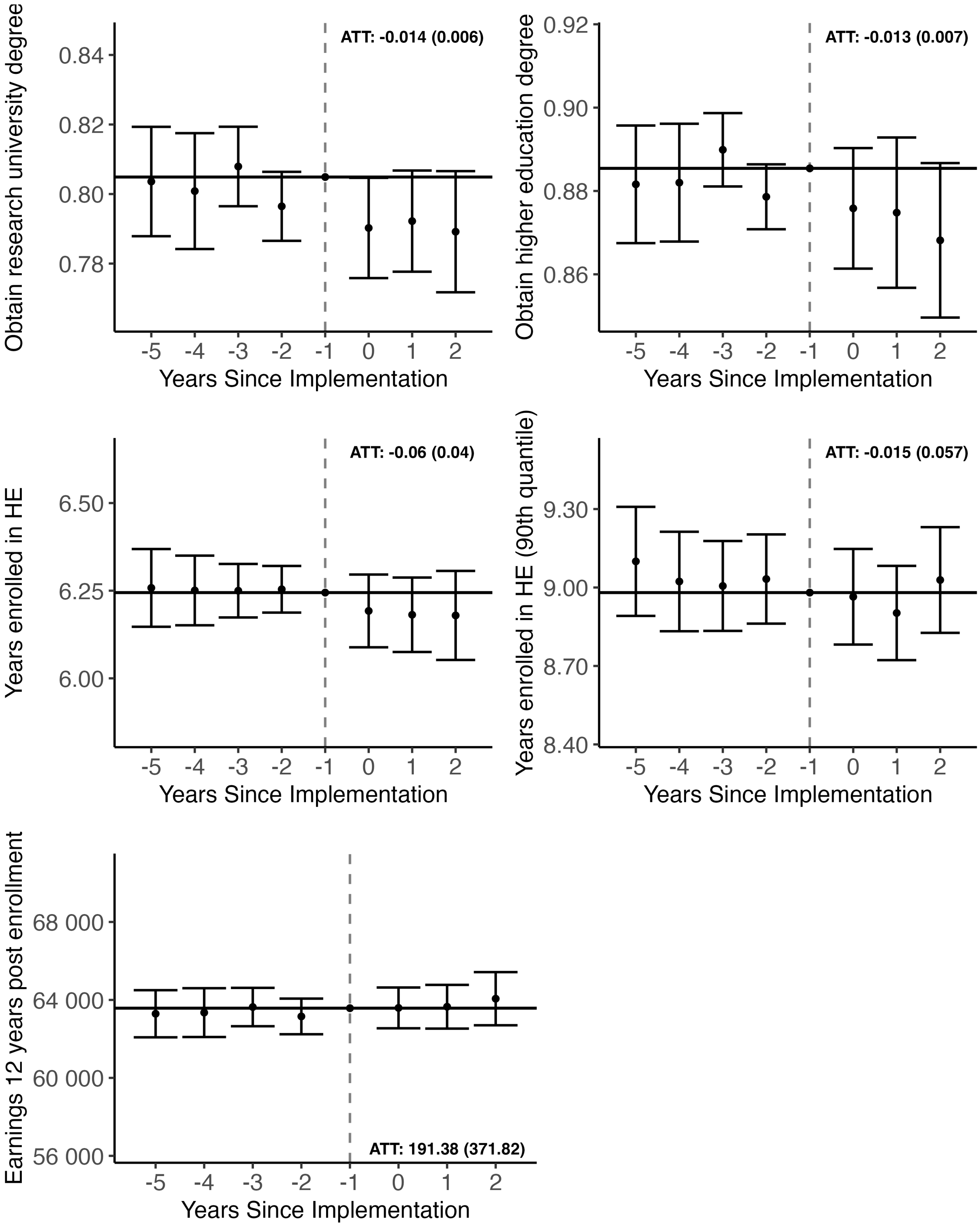
Note: This figure presents the event-study estimates corresponding to the results in Table 3. The graphs are centered around the average of the outcome in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing. RU stands for Research University. UAS stands for University of Applied Sciences.

Figure A5: Event study estimates from Table 4



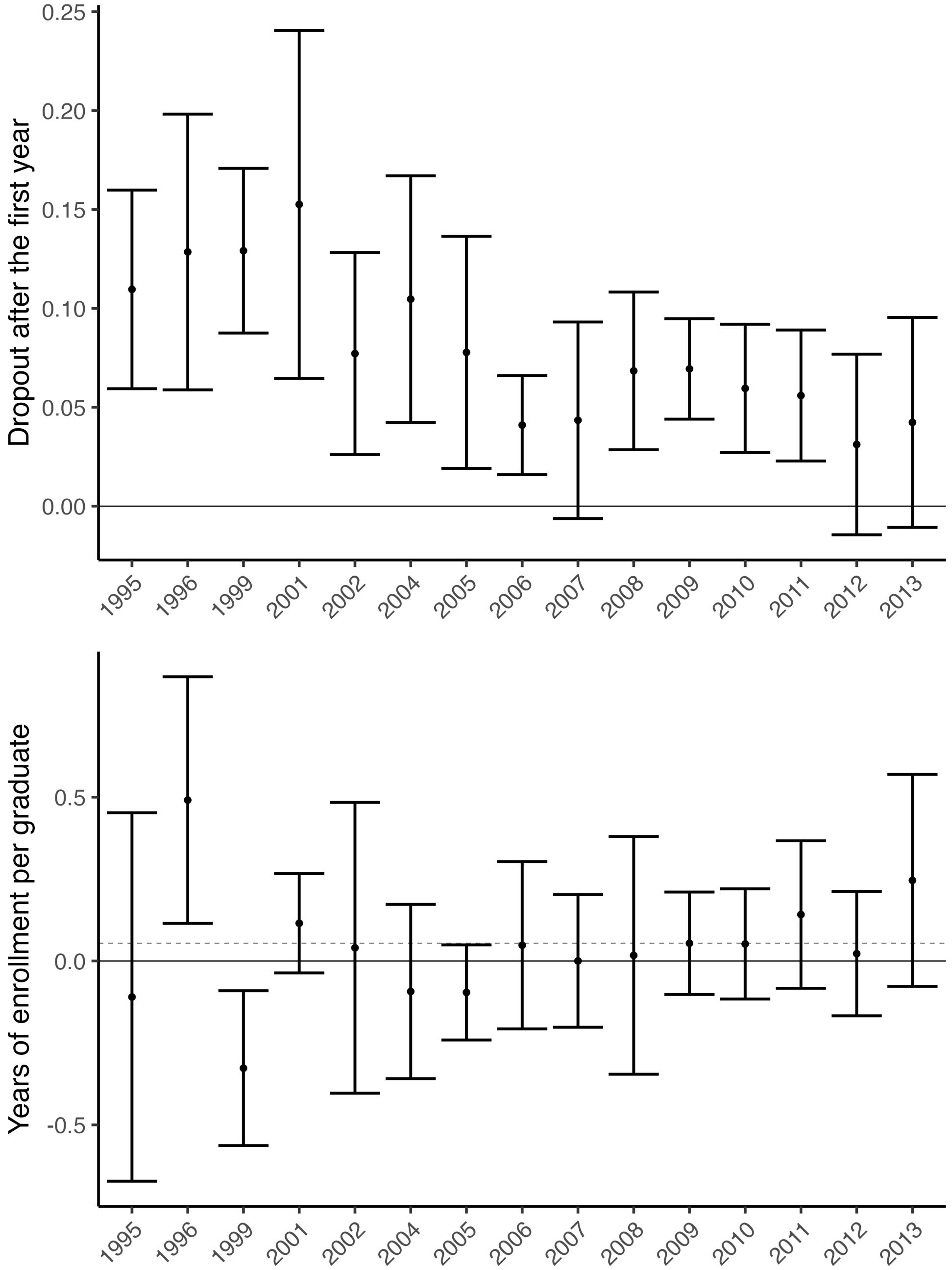
Note: This figure presents the event-study estimates corresponding to the results in Table 4. The graphs are centered around the average of the outcome in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing.

Figure A6: Event study estimates from Table 5



Note: This figure presents the event-study estimates corresponding to the results in Table 5. The graphs are centered around the average of the outcome in the last pre-treatment period. The error bars present 95% confidence intervals, adjusted for multiple hypothesis testing. HE stands for higher education.

Figure A7: Effects of implementing performance standards across adoption cohorts



Note: this figure presents estimates of the effect of implementing performance standards on first-year dropout (upper panel) and the average enrollment duration in higher education per graduate (lower panel) for all adoption cohorts. The estimates are constructed using a group-level decomposition of the full sample ATT. This methodology is explained in further detail in Callaway and Sant'Anna (2021). The dashed line is centered around the ATT for the full sample. The error bars present 95% confidence intervals.

Appendix B: Predicting Student Success

In Table 2 and Figure 4, I make use of students' predicted field-specific on-time graduation chances. This section explains how these predictions are constructed.

I employ gradient-boosted decision trees to predict students' chances of graduating within 4 years from their chosen program (Friedman (2001)). The explanatory variables include: gender, age, the highest income and education of the parents, migration generation, previously enrolled at a university of applied sciences (HBO), the level of high-school education, the field of study, and the subfield of study. The field of study and the subfield of study are based on classifications from the Ministry of Education. Missing values are replaced by indicators rather than excluded from the analysis. To ensure the independence of my predictions from the analysis sample, I train the models using only students who neither serve as the treatment nor control group. These consist of students in programs that implemented the BSA at least four years ago or will do so at least five years later.

Table B1 reports the performance of the prediction exercise using students from the core analysis sample who are not used for training the prediction model. Students are divided into four equally sized groups, ranging from the lowest to the highest predicted probability of on-time graduation. Columns 2 and 3 show that the predictions are accurate on average. The remaining columns indicate that students with low predicted probabilities are more often male, relatively old, and have less educated parents. The final column reports secondary school GPA, which is available only for cohorts since 2007. Although grades were not used in the prediction model, this column shows that students with low predicted probabilities also have lower grades. This confirms that the model accurately distinguishes between students of lower and higher ability.

Table B1: Prediction results in the unseen analysis sample

Prediction quantile	Prediction ($\hat{y}$)	Graduation within four years (y)	Male	Age	Income parents	Education parents	Migration Generation	Previously enrolled in HBO	Secondary school GPA
25	0.28	0.28	0.80	20.51	0.74	15.22	0.12	0.17	-0.24
50	0.38	0.39	0.61	19.36	0.75	15.6	0.10	0.11	0.12
75	0.44	0.44	0.40	19.01	0.76	15.62	0.12	0.08	0.24
100	0.51	0.50	0.13	18.94	0.76	15.59	0.16	0.07	0.37

Notes: This table reports results from the trained machine learning model on the unseen analysis sample. The machine learning model used is a gradient boosted decision tree (XGBoost). The model is trained to predict the likelihood that a student graduates within four years from the initial program. The training data includes all students from the core sample who were enrolled in programs at least 5 years before or 4 years after this program implemented a performance standard. The model is tuned using five-fold cross-validation. I sort the samples into predicted graduation quartiles (column 1) and calculate their predicted (column 2) and realized graduation rates (column 3). The remaining columns show averages across the different quartile.

Appendix C: Deterrence Effects

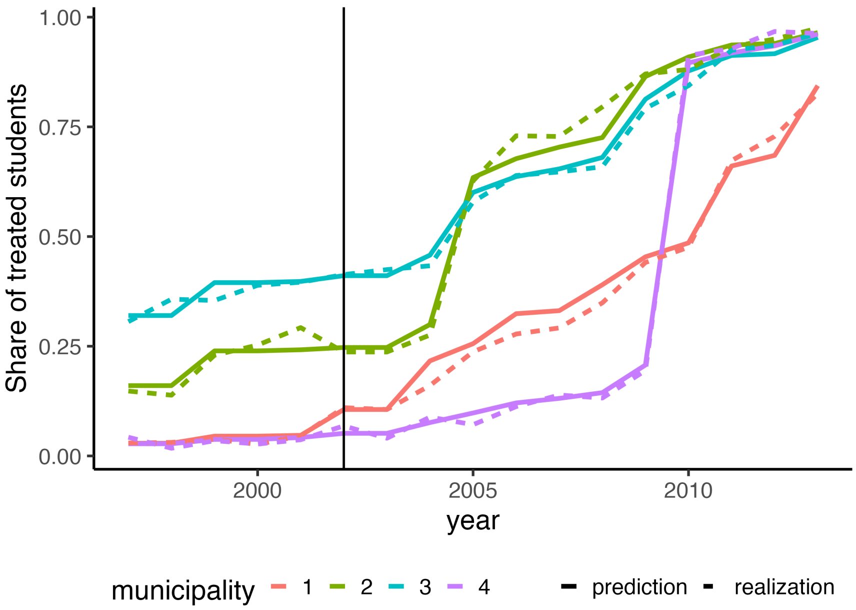
This appendix tests for deterrence effects using an alternative identification strategy. I show that prospective students who lived in the same municipality at age 16 choose similar programs over time, and that this pattern is unaffected by the introduction of performance standards. This suggests that performance standards neither deter university enrollment nor induce students to shift to other programs.

Specifically, I first select all students from the core sample for whom their municipality at age 16 is known (86.5 percent). For each municipality, I then calculate the share of students from this municipality that enrolled in each program between 1993 and 2002. For example, 1 percent of students from a given municipality could be enrolled in economics at the Erasmus University Rotterdam, 2 percent in mathematics at Leiden University, and so on. Based on these pre-2003 shares and the adoption rate of performance standards in these programs, I predict for each municipality how many students are expected to be in a program with a performance standard for each year after 2002. If students do not change their enrollment behavior, then these predicted shares should correspond to the observed shares.

To illustrate this approach, Figure C1 shows the predicted and realized share of students exposed to a performance standard for the four municipalities with the largest number of individuals who lived there at age 16 and later enrolled at a university. Consider municipality 2. Between 2002 and 2009, predicted exposure increases by about 20 percent, whereas in other municipalities, it rises more quickly. This is because only a few previously popular programs in municipality 2 implemented a performance standard during this period. In 2009, however, the predicted exposure jumps to almost 90 percent because many previously popular programs among students from this municipality implemented a performance standard that year. These programs are from a nearby university that adopted the performance standard in 2009. Crucially, the figure shows that the observed share of treated students closely mirrors the predicted share.

This also holds for the broader sample. To show this, I regress a student's observed treatment status after 2002 on their predicted treatment status based on the pre-2002 shares, including municipality and year fixed effects. These fixed effects ensure that variations in the predicted treatment are due to the staggered implementation of performance standards by previously popular programs. The coefficient equals 1.004 (Table C1), indicating a virtually perfect alignment between predicted and realized shares. This shows that whenever previously popular programs implemented a performance standard, the number of students exposed to these standards increased proportionally. This is inconsistent with deterrence effects.

Figure C1: Predicted and Realized Share of Treated Students in Four Municipalities



Note: the dashed lines show predicted shares of students enrolled in programs with a performance standard for the four municipalities with the largest student population. Predictions are constructed as follows: (i) select all first year students between 1993 and 2002 who lived in the municipality at age 16, (ii) calculate the share of these students in each program, and (iii) for a given year, sum the shares of students in programs that have a performance standard in that year. The solid lines show the observed shares. The vertical line at 2002 marks the last year used to construct predictions; all predicted shares after 2002 are therefore based solely on pre-2003 enrollments and evaluated against observations not used in generating the predictions.

Table C1: Testing for deterrence effects at the municipality level

	Observed treatment status
Predicted treatment status	1.004***
	(0.014)
N	306,977
Municipality fixed effects	x
Year fixed effects	x

Notes: This table reports results from a regression of an indicator for enrollment in a program with a performance standard on the student's predicted likelihood of such enrollment. The specification includes fixed effects for municipality at age 16 and year of enrollment. Predicted treatment status is constructed as follows: (i) select all first year students between 1993 and 2002 who lived in the same municipality at age 16, (ii) calculate the share of these students in each program, and (iii) for a given year, sum the shares of students in programs that have a performance standard in that year. The estimation sample consists of all first year students in the core sample who enrolled after 2002 and for whom municipality at age 16 is observed. Restricting to post 2002 cohorts ensures that predictions are evaluated on observations not used to construct them. Standard errors are in parentheses. (*** $p < 0.01$,$^{**}p < 0.05$,$^{*}p < 0.1$)

Appendix D: Survey evidence

This section describes the data collection, phrasing of questions, and the sample I use for the survey experiment.

The data come from a survey administered to first-year bachelor students from the public university Vrije Universiteit (VU) Amsterdam over a one-week period in November 2024. The students were recruited from a research participation course run by the behavioral lab of the School of Business and Economics. Students were informed that the study consisted of several questions related to their expectations and performance in the first year of the program.

The data were collected through an online survey (using Qualtrix). The survey took approximately 15 minutes to complete and consisted of several parts. Students received a small number of course credits upon the completion of the survey. I also included two attention checks and informed students that failing this results in withdrawal from the study. Of the 343 respondents, 10 failed the attention check, resulting in an overall sample of 333 students.

Subjective expectations. In addition to questions about demographics, personality traits, and secondary school performance, there are two main questions used for the analysis. The first question is aimed at eliciting students' expected number of course credits by the end of the first year. The specific question is:

At this moment, how many credits do you expect to have by the end of the first year? Divide a 100 points over the following options:

(30 or less, 31-36, 37-42, 43-48, 49-54, 55-60)

For instance, assigning 20 points to 31-36 credits implies a 20% chance of attaining between 31 and 36 credits by the end of the year. Assigning zero points signifies a belief that obtaining that specific number of credits is not possible.

The question has a built-in logical check, so that the percent chances of the number of course credits always sum to 100.

The performance threshold in these bachelor programs is 48 credits, and all courses yield 6 credits upon passing. I construct the subjective probability of dismissal by summing the probabilities of receiving 30 credits or less, between 31 and 36 credits, and between 37 and 42 credits.

Willingness to forgo gifts. The second set of questions present the hypothetical choice scenarios. In order to maximize respondents' reliability, I first present the following explainer:

We will present you some hypothetical situations where you have to choose between two alternatives. It is important that your choices here are realistic. Hence, in each of the following choice tasks, please consider carefully that you would actually prefer the alternative you choose.

After this, the main question reads:

Suppose you were given the choice between the following: receiving a payment today or an immediate removal of the Binding Study Advice. This would imply that you may continue with the program next year, regardless of how many course credits you have by the end of the year. We will now present to you five situations. The payment is different in every situation. For each of these situations we would like to know which you would choose.

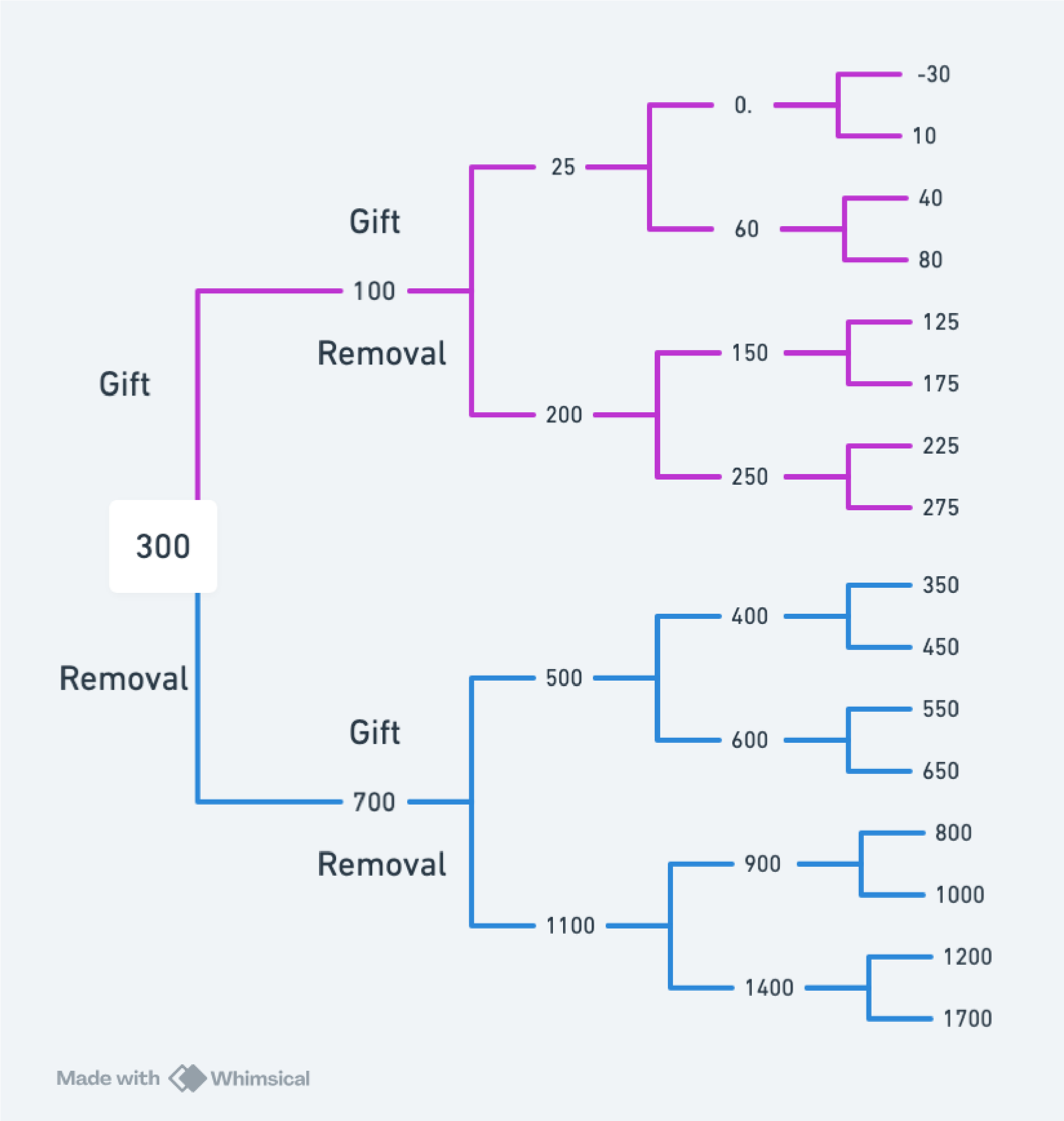
Each of the following five questions is of the following form:

Would you rather receive a gift of X Euro today or an immediate removal of the Binding Study Advice?

The first question always corresponds to $X = 300$. Whenever a student chooses 'Gift', then the next question decreases the gift value, and if the student chooses for the removal of the BSA, then the next question increases the gift value. The values of X vary according to the scheme in Figure D1.

There are two special cases, indicated by $X = 0$ and $X = -30$ in Figure D1. If students choose a gift of 25 euros over the removal of the performance standard, then they get the following question (denoted by $X = 0$ in Figure D1): *What would you prefer: '**keeping** the Binding Study Advice (BSA) over its immediate removal?'* If a student chooses 'Keeping the BSA', then the next question is (denoted by $X = -30$ in Figure D1): *Would you prefer **paying** 30 euro today or an immediate removal of the Binding Study Advice?* If the student preferred the removal of the BSA over keeping the BSA, then the next question corresponds to $X = 10$.

Figure D1: Staircase method scheme


-30
0.
10
25
40
Gift
60
80
100
Removal
125
150
Gift
175
200
225
250
275
300
350
400
450
Removal
500
550
Gift
600
650
700
800
Removal
900
1000
1100
1200
1400
1700
Made with Whimsical