Performance Standards in Education

Sander de Vries*

February 21, 2025

Abstract

Although many education programs use performance standards to increase student performance and dismiss unfit students, their effects are unclear. This paper studies the effects of implementing strict performance standards in the first year of bachelor programs in the Netherlands. Using a difference-in-differences design and rich administrative data, I show that performance standards do not deter prospective students from enrolling, even though they considerably increase the risk of dismissal. In the long run, implementing performance standards does not improve students' education degree attainment, enrollment duration, or subsequent labor outcomes. Additional survey evidence suggests that students experience considerable disutility due to performance standards.

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

JEL Codes: I23, I21, J16

*Department of Economics, Vrije Universiteit Amsterdam, s.de.vries@vu.nl. I gratefully acknowledge valuable comments from Nadine Ketel, Maarten Lindeboom, Hessel Oosterbeek, and conference and seminar participants in Amsterdam, The Hague, Rotterdam, Leiden, Trier, and London. 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

Many tertiary education programs have academic dismissal policies designed to remove students who fail to meet specific performance standards. For instance, many colleges and universities in the US, Canada, Australia, New Zealand, Mexico, South Korea, Singapore, and India implement probation-dismissal policies. These policies place low-performing students on probation and warn them that they will be dismissed if they fail to improve in subsequent terms. Other performance standards include limiting the number of exam retakes (as in Germany, Switzerland, and Austria) or requiring students to earn a certain number of credits during the initial study phase (as in the Netherlands, Indonesia, and South Africa).[1] These rules aim to address low graduation rates and improve student performance. Despite their prevalence in education, we know little about the effects of performance standards.

Theoretically, effects are uncertain. Positively, they can motivate students to exert greater effort in their initial years, potentially leading to better academic outcomes overall. They can also help redirect unfit students toward more suitable career paths, contributing to higher graduation rates, faster completion times, or greater productivity. However, performance standards can have adverse effects when they result in the dismissal of students who would have otherwise graduated, if dismissed students lack better alternative career paths, or when they diminish students' intrinsic motivation. Moreover, performance standards can deter prospective students from enrolling, which can work out positively if these are unfit students, but it may also impose barriers to suitable students.

Given these contrasting outcomes, it is important to evaluate the implications of performance standards empirically. To this end, I study the effects of performance standards in the first year of Dutch university programs. These performance standards require students to achieve a specific threshold of course credits — typically around 65 percent - by the end of their first year.[2] Unless students are exempted for personal reasons, failing to meet this

[1] This is no exhaustive list. Nearly every country has mechanisms to dismiss poor-performing students, although the strictness of such policies differs greatly between countries.

[2] Officially, these programs provide all students a verdict at the end of the first year, indicating whether

standard results in students being removed from the program.

The Netherlands provides an excellent setting for studying performance standards. The fraction of programs with performance standards increased gradually from no to nearly all programs between 1994 and 2014, enabling me to use a staggered difference-in-differences design to evaluate its effects. As the last programs implemented the performance standard over ten years ago, I can also study the long-term effects on degree completion and labor market outcomes. These are crucial outcomes, as the goal of the performance standards was to improve students' graduation rates or time-to-graduation. Moreover, the performance threshold used by bachelor programs is relatively high: on average, one of every five students fails the performance standard. This means that its effects are not limited to marginal students at the very bottom of the performance distribution.

Despite the widespread use of performance standards in higher education, there is a notable gap in the literature concerning their implications. Most existing studies use regression discontinuity designs to study the consequences of *being on probation* or *being dismissed* due to failing to meet these standards (Lindo et al. (2010), Fletcher and Tokmouline (2018), Ost et al. (2018), Wright (2020), Albert and Wozny (2022)). The key contribution of this study is to provide unique evidence on the broader effects of *implementing* performance standards.[3] This approach is important as the introduction of performance standards can potentially alter the selection of students into academic programs and prompt shifts in the entire performance distribution. Moreover, marginal students may be differentially affected by the policy. Such mechanisms remain unexplored with regression discontinuity designs.

The second key contribution comes from using rich administrative data from Statistics Netherlands that enables me to observe all students' career trajectories up to twelve years

they can continue. This verdict is known as the Binding Study Advice (BSA). The specific performance standards differ between faculties and universities but are often close to 65 percent. The details of the policy are discussed in section (2).

[3]This distinction is similar to the literature on high-stakes exit exams, where some papers use RD 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. See, for example, Ou (2010), Clark and Martorell (2014), Caves and Balestra (2018), Bach and Fischer (2020), ter Meulen (2023), Fidjeland (2023) and references therein.

after enrollment. Unlike most of the literature, these data also include the trajectories of dismissed students.[4] This is especially valuable because the welfare effects of academic dismissal policies hinge on the ability of dismissed students to secure alternative career paths.

This paper also offers new insights into two broader questions in education research. First, it speaks to the literature on the optimal design of performance incentives in education. Many papers consider relatively small-scale interventions, short-run outcomes, or positive incentives.[5] This paper provides unique evidence on the *long-run* effects of a *large scale* implementation of a *negative* performance incentive.

Second, this paper provides rare evidence of how prospective students respond to exogenous changes in programs' performance criteria. Arpita et al. (2021) summarize the recent literature on major choice and conclude that major choice is mostly determined by major-specific 'tastes', rather than earnings differentials. In an attempt to understand these tastes better, previous studies mostly focus on demand-side explanations such as preferences for course enjoyability, parental approval, peer effects, work-family balance, and marriage market considerations (Zafar (2013), Wiswall and Zafar (2018), Wiswall and Zafar (2021), Altmejd et al. (2021)). However, little is known about how changes in the programs' characteristics, such as difficulty or grading, affect students' choices. In theory, one would expect that some prospective students are deterred by a performance standard because they do not want to raise their effort levels or risk a dismissal.

On the contrary, I find no evidence of deterrence effects. The introduction of a performance standard does not affect the number of enrollments on average, nor for specific adoption cohorts or fields of study. The composition of new students in terms of gender, socioeconomic background, or student quality is also unaffected.

The lack of a deterrence effect is especially interesting given that the performance stan-

[4]To the best of my knowledge, Ost et al. (2018) is the only study on academic dismissal policies that contains outcomes of dismissed students after they drop out.

[5]Examples include papers that study different mechanisms to award financial rewards (Angrist and Lavy (2009), Rodríguez-Planas (2012)), combinations of financial incentives and support services (Angrist et al. (2009)), goal-setting strategies (Clark et al. (2020)), or the effects of receiving different 'labels' or grades for high-stakes exams (Papay, Hvidman and Sievertsen (2021)).

dards substantially affect students' trajectories. Although I do not observe the actual dismissal status, I observe that the number of students that drop out after the first year increases by 7 percentage points (25 percent). This suggests that a substantial number of students who would otherwise have continued are dismissed. Most students re-enroll in other university programs, while about 15 percent re-enroll in Universities of Applied Sciences (colleges), and 20 percent leave higher education. Moreover, 70% of dismissed students switch institutions.

Although the government introduced the performance standards to increase the match quality between students and programs, I find no evidence that students nor the government benefit from the dismissal policy. Performance standards decrease the time that students are enrolled in their initial program, but their overall enrollment duration in higher education is not affected. Moreover, students are equally likely to obtain a degree from higher education as they would have been in the absence of the performance standard. There are also no effects on labor market outcomes twelve years after the initial enrollment.

I also explore adverse effects on student well-being. I find no evidence that the performance standard increases the usage of psychostimulants or antidepressants. However, using additional survey evidence, I find suggestive evidence that the presence of a performance standard decreases students' utility. I try to gauge the extent of (dis)utility by asking first-year students hypothetical choice questions about how much money they are willing to forgo for immediate removal of the performance standard. More than 80 percent of students are willing to forgo money, with the median amount being 412 euros, and almost 30 percent indicating the maximum amount of 1100 euros.[6] Even students with low subjective risks of dismissal are willing to forgo money.

Overall, my results provide a cautionary tale for the effectiveness of performance standards in education. Although they can have substantial effects on students' career trajectories, they need not improve long-term education or labor market outcomes, and their presence appears to induce substantial disutility among students.

[6]The average monthly disposable income of students is estimated to be 943 euros.

2 Background and Institutional Context

2.1 Literature

Regression discontinuity (RD) evidence. Most studies examining the effects of performance standards use RD designs, comparing students performing just above the standard to those just below it. This approach was first employed by Lindo et al. (2010), who find that being on probation at a Canadian university both discourages some students, leading to attrition, and motivates others, resulting in short-term performance gains, though there are no long-term effects on graduation rates. Fletcher and Tokmouline (2018) identify comparable effects at four Texan universities.

In contrast, Albert and Wozny (2024) find that academic probation at the U.S. Air Force Academy led to higher STEM degree completion rates without increasing attrition. They attribute this to the combination of probation with an intensive support program. This is consistent with Canaan et al. (2022), who find that targeted coaching for students on probation significantly improved academic performance, especially among low-income students.

Probation also appears to influence students' course selection. Casey et al. (2018) and Wright (2020) indicate that students on probation often enroll in fewer and easier courses, suggesting a strategic response to academic pressure.

Finally, instead of focusing on the effects of being placed on probation, Ost et al. (2018) investigates the consequences of academic dismissal at 13 universities in Ohio. By comparing those who perform just above or below the dismissal threshold, they find significant long-term earnings losses for students who are dismissed.

Evidence from the Netherlands. Two closely related studies from the Netherlands also explore performance standards. Arnold (2015) uses a fixed effects design focussing on bachelor programs between 2002 and 2007 to find that performance standards increase first-year dropout rates but improve four-year completion rates among those who remain. Similarly,

Sneyers and De Witte (2017) use a two-way fixed effects design (2003-2008) and find increased dropout rates and improved graduation rates but decreased student satisfaction.

Despite their valuable insights, these studies face notable data limitations. Both rely on aggregated program-level outcomes instead of individual-level data. Arnold (2015) is constrained to analyzing four-year graduation rates without the ability to pinpoint the exact timing of graduation. More critically, Sneyers and De Witte (2017) do not observe actual degree-level outcomes; instead, they impute these outcomes using aggregated field-of-study data. They are also unable to track the trajectories of dismissed students.

This paper addresses several significant gaps in the literature on performance standards. Its primary contribution is to provide novel evidence on the effects of implementing performance standards, as opposed to solely examining the consequences of probation or dismissal. To the best of my knowledge, existing Dutch studies are the only other attempts to investigate this aspect. This paper extends these studies by utilizing individual-level data with granular enrollment information covering *all* programs that implemented performance standards. Additionally, it explores how these standards influence students' program choices, subsequent labor market outcomes, and wellbeing, offering a comprehensive view of their broader impacts.

2.2 Higher Education in the Netherlands

Higher education in the Netherlands is provided by Universities and University of Applied Sciences (colleges). This paper focusses on bachelor programs offered at universities. These bachelor programs offer specialized undergraduate education throughout the entire program, unlike, for example, in the United States, where students specialize later. The bachelor programs are offered by thirteen universities, with three specializing in technical sciences and one in agricultural science, while the others offer a broad spectrum of bachelor's education across various disciplines.

Admission to these programs primarily requires a high school diploma from the prepara-

tory academic track or completion of the first-year curriculum at a University of Applied Sciences.[7] Students from international backgrounds must present qualifications equivalent to the Dutch preparatory academic track to gain admission.

Dutch bachelor programs use the European Credit Transfer System (ECTS). 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. As a result, the median time to completion is 4 years instead of 3. Once students accumulate the required 180 credits, they are granted a bachelor's degree and become eligible to pursue a related master's program.[8]

Dutch university education possesses two distinctive features. First, the majority of bachelor programs are required to admit all 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. 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.[9]

[7]Students are categorized into different tracks upon entering high school at the age of 12, with only the highest level, the preparatory academic track, ensuring eligibility for university education. Some bachelor programs, primarily in technical fields, may necessitate the completion of specific prerequisite courses from their prior education.

[8]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, the majority of the programs introduced the performance standard after 2003 and therefore my results mostly apply to bachelor programs (rather than undivided programs). I find similar results when I exclude the years before 2003.

[9]To support this point empirically, Avdeev et al. (2023) 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.

2.3 The Performance Standard at Dutch Universities

In the early 1990s, Dutch universities struggled with high dropout rates and students taking an extended time to graduate. This was particularly salient among students who underperformed in their first year. To address these issues, universities were mandated in 1993 to offer advice to first-year students on whether to continue in their chosen academic programs. The intent was to redirect students who were unlikely to succeed into more suitable fields of study. Importantly, universities had the discretion to decide whether this advice would be mandatory or merely advisory.[10] When enforced as mandatory, this guidance is known as the Binding Study Advice (BSA). Students receiving a negative BSA are dismissed from their program and prohibited from re-enrolling in the following year. They are allowed to enroll in the same program at another university or in other programs at the same university.[11]

When universities enforce the Binding Study Advice (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 decisions, and students have the right to contest negative advice.

The option to implement a performance standard sparked significant debate among university and faculty boards, which had the authority to decide whether to adopt the Binding Study Advice (BSA) at their institutions. On one hand, the Dutch government pressured universities to shorten the time students spend in programs and improve graduation rates,

[10] This regulatory framework is specified in the in the Higher Education and Scientific Research Act (in Dutch: "WHW") of 1993

[11] At some universities, students are also not allowed to enroll in programs in the same study field.

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.[12] On the other hand, many academics and students strongly opposed dismissing students based solely on their first-year performance.[13]. Critics also argued that the BSA would add administrative burdens, including an increase in appeals and procedural complexities.

Due to differing attitudes toward the BSA, its implementation varied widely among universities, with the first full adoption occurring in 1996. The pace of adoption increased in the mid-2000s as the government intensified pressure on institutions to improve academic efficiency. By 2014, nearly all university-level bachelor programs had adopted the BSA.

3 Data

3.1 Implementation of the Performance Standard

I collected data on the introduction of the BSA at programs from multiple sources. The Association of Dutch Universities provided me with records of programs between 1995 and 2012. For the subsequent period, I leveraged data from Bachelors.nl, a website tailored for prospective students, featuring comprehensive information about university bachelor programs. Through a web archive, I accessed records dating back to 2011. To ensure data accuracy and reliability, I cross-referenced all records with the annual reports from programs, faculties, or universities whenever accessible.

Figure 1 illustrates the progression of BSA adoption among programs. The rollout is staggered, with additional programs incorporating the BSA each year since 2003. Notably, programs rarely implemented the BSA in isolation. Instead, a faculty or the entire university typically implemented the BSA across all programs simultaneously. This approach strength-

[12]See "Serieus werk maken van bindend studieadvies", Volkskrant (2006).

[13]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)

ens the argument that the timing of BSA implementation is likely independent of trends in specific programs. In the main results section, I show evidence to support this claim.

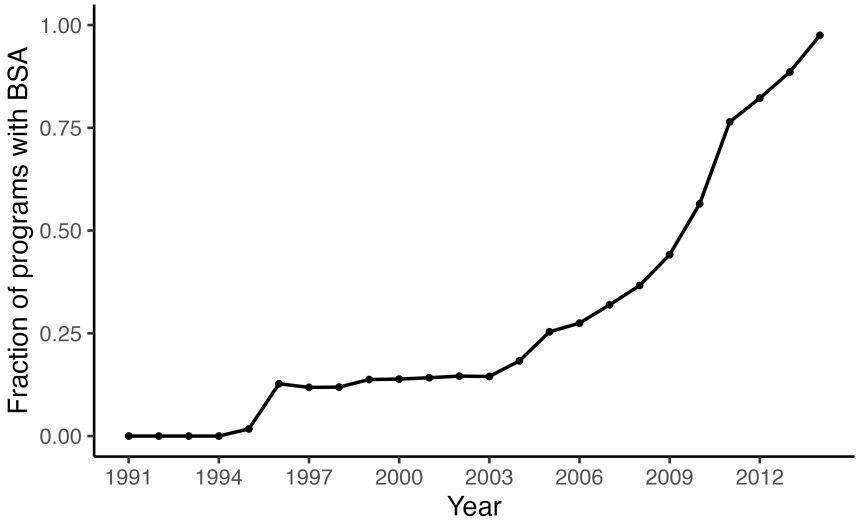


Figure 1: Staggered rollout of the performance standard in bachelor programs

3.2 Sample selection and administrative data

I use administrative data on the universe of students at Dutch universities from Statistics Netherlands. All names have been replaced with anonymous identifiers. These identifiers allow researchers with approved access to join records associated with an individual across a range of government services, including the personal register, enrollments in education, graduation records, wages, and medicine prescriptions. Below, I discuss sample selection and key variables, and provide descriptive statistics.

Sample. The enrollment records encompass data for all bachelor program enrollments at universities dating back to 1991. I include all students who first enrolled at a university between 1991 and 2014, with each observation representing 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. Moreover,

I exclude students who were previously enrolled in a university of applied sciences for more than two years, because it is unclear from the data whether these students participate in first-year courses. Following these selection criteria, my sample covers roughly 700.000 students across 362 distinct bachelor programs.

Key variables. First, I construct an indicator that equals one when students do not continue their studies after the first year. As I do not observe the actual BSA verdict, a dropout can be voluntary or because of dismissal. I also analyze what students do after dropping out. In particular, I distinguish between four alternatives: (i) re-enroll in a university program in the same field, (ii) re-enroll in a university program in another field, (iii) enroll at a university of applied sciences, or (iv) all else. The study field is based on a classification of programs into 9 fields by the Ministry of Education in the Netherlands, which includes fields such as engineering, sciences, economics, and law.

Next, I measure the number of years a student is enrolled in the initially chosen program or in higher education in general. While the results are presented in years for ease of interpretation, the underlying enrollment data is recorded in months, affording a granular level of analysis.[14] The total enrollment duration in higher education is counted from the start of the initial enrollment until the end of the last enrollment. This also includes breaks like gap-years. The results are the same when I define the enrollment duration as the sum of the yearly enrollments.[15]

I similarly derive two graduation indicators. The first is an indicator of whether a first-year student ultimately graduates from the initially chosen program. The second is an indicator of whether the student attains a bachelor's degree from any program, potentially different from their initial choice.[16]

[14]Students receive 1/12th of the tuition fee back for each month they have stopped before the end of the academic year. Short enrollments are quite common in the data as a result. Moreover, the enrollment data are based on tuition payments rather than graduation certificates. This allows me to observe the study duration for all students, including those who never completed a program.

[15]Results are available upon request.

[16]The latest cohort in my analysis starts in 2013. Since I have graduation data until 9 years later, the data is not censored for most of these students. Yet, there may be a very small number of students who I label as not-graduated even though they eventually will.

The income data include all earnings from employment and entrepreneurship up to 2023. All incomes are adjusted to 2015 prices. Labor market participation is defined as having any non-zero earnings. I examine earnings 12 years after the initial enrollment, which corresponds to an average age of 31. Income information is only available for cohorts that enrolled before 2012.

Lastly, to assess mental health effects, I use data on prescriptions filled at pharmacies and covered by basic health insurance.[17] This dataset encompasses information regarding whether a student received at least one prescription for antidepressants (ATC-4 code: N06A) or psychostimulants (ATC-4 code: X) in a given calendar year. To bolster statistical power, I aggregate medication prescriptions up to five years following the start of the first academic year. As medicine prescriptions are only available from 2007 onward, I restrict the sample to programs that implement a BSA after 2006 for this outcome.

I supplement these data with information about demographics, secondary school centralized exam grades, and parental income. These variables are used to test whether implementing a performance standard affects the composition of newly enrolled students and for heterogeneity analysis. Table 1 reports the descriptive statistics of students' characteristics and outcomes.

3.3 Predicting students' graduation chances

In the analysis, I make use of students' predicted program-specific graduation chances. This is useful for two reasons. First, these predictions enable me to test whether imposing a performance standard discourages students with low chances of success in that specific program. Second, the predicted graduation chances offer an interesting dimension for heterogeneity analysis. The primary effects of the performance standard are expected to be at the bottom of the performance distribution. The predicted graduation probabilities provide an intuitive way to zoom in on students who are likely to perform poorly.

[17] In the Netherlands, since 2006, all residents are required to purchase health insurance with a predefined set of benefits from private insurers. Most antidepressants are covered by this basic insurance.

I employ gradient-boosted decision trees (Friedman (2001)) to predict students' chances of graduating in their program. I include an extensive range of variables, including demographics, parental income and education, nationality, prior education, migration background, and more. For the years 2007-2013, I also have detailed data on students' secondary school grades. As these are potentially valuable predictors of students' future performance, I also train a separate model that includes these variables. To ensure the independence of my predictions from the analysis sample, I train the model using only students who neither serve as the treatment nor control group. These consist of students in programs that implemented the BSA over four years ago or will do so at least four years later.

Table 2 reports the performance of the prediction exercise in the *unseen* analysis sample, as well as some characteristics of students with high or low predicted graduation probabilities. Columns 2 and 3 show that the predictions are accurate for both models. Interestingly, panel B shows that the model including detailed course-specific secondary school grades only marginally refines the predictions. In both models, even the riskiest 25 percent of students maintain a graduation likelihood of approximately 50 percent. This suggests that, with or without grades, it is very difficult for programs to accurately classify future dropouts based on pre-enrollment information. For my analysis, precise classification is unnecessary; estimation of prediction probabilities suffices.

4 Identification

My analysis of the implementation of performance standards in bachelor programs is fairly straightforward. First, I collapse all the data to the program level. Whenever I refer to a program, I refer to a specific bachelor's program at a specific institution. For each program p in year t, I then observe an outcome Y_{pt}. Unless otherwise stated, the outcomes are averages taken over all students in my sample who are initially enrolled in program p at time t. In some cases, I will also study specific parts of the outcome distribution, such as the lowest

Table 1: Descriptive Statistics

	Cohorts	Mean	SD	N
Characteristics				
Male	All	0.49	0.5	703,463
Age	All	19.38	1.49	703,463
Foreign	All	0.12	0.32	703,463
previously enrolled in HBO	All	0.06	0.24	703,463
Income Percentile Father	All	0.69	0.27	588,480
Centralized Secondary School Exam GPA	2006-2014	6.49	0.79	234,343
Decision after the first year				
Continue	All	0.75	0.43	703,463
Switch to the same field	All	0.06	0.24	703,463
Switch to another field	All	0.07	0.26	703,463
Switch to HBO	All	0.06	0.25	703,463
Other	All	0.05	0.22	703,463
Enrollment duration and graduation outcomes				
Graduate from initially enrolled program	All	0.62	0.49	703,463
Graduate from any program	All	0.87	0.33	703,463
Number of years enrolled in the first program	All	3.73	2.48	703,463
Number of years enrolled in higher education	All	6.86	3.45	703,463
Labor outcomes				
Non-negative earnings (12 years after enrollment)	1991-2010	0.96	0.19	476,920
Income (12 years after enrollment)	1991-2010	53304	33300	476,920
Medicine usage				
Used antidepressants within five years after enrollment	2007-2014	0.05	0.22	310,999
Used stimuli within five years after enrollment	2007-2014	0.03	0.18	310,923
Program size	All	90.86	116.67	362

Notes: this table presents descriptive statistics for all variables considered in the analysis.

Table 2: Prediction results in the unseen analysis sample

Prediction quantile	Prediction ($\hat{y}$)	Graduation (y)	Male	Age	Income parents	Migration Generation	Pre-HBO	Grade	N
	A. Full-sample (without grades)								
$0-25$	0.52	0.52	0.82	20.18	0.63	0.13	0.09	6.25	$50,286$
$26-50$	0.64	0.64	0.61	19	0.68	0.12	0.05	6.52	$50,281$
$51-75$	0.7	0.71	0.38	18.99	0.68	0.17	0.05	6.54	$50,274$
$76-100$	0.79	0.79	0.11	18.75	0.71	0.11	0.04	6.61	$50,306$
	B. Cohort 2007-2014 (including grades)								
$0-25$	0.46	0.48	0.83	19.9	0.65	0.13	0.06	5.84	$34,754$
$26-50$	0.63	0.63	0.53	19.29	0.66	0.18	0.05	6.16	$34,754$
$51-75$	0.73	0.73	0.33	19.05	0.67	0.19	0.05	6.54	$34,754$
$76-100$	0.84	0.84	0.32	18.63	0.71	0.04	0.03	7.26	$34,755$

Notes: I predict graduation from a program using an Extreme Gradient Boosting algorithm trained on students who were enrolled in programs at least 4 years before or after this program implemented a performance standard. The model is tuned using five-fold cross-validation. Panel A presents the predictive performance of the machine learning model that predicts graduation for the full sample but without information about grades. Panel B presents the prediction results of the model which includes grades (only available after 2007). 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 descriptive statistics of the different quartiles.

grades or longest enrollments.

As a baseline, I estimate event-study regressions of the form

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

where τ_{pt} is the years relative to treatment, with $\tau_{pt} = 0$ in the first treatment period. Moreover, α_p and γ_t are program and year fixed effects. The regressions are weighted by program size. 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.

Assumptions. The main assumption is that - in the absence of the treatment - the outcomes of students in treated programs would have evolved in parallel to that of students in untreated programs. There are multiple reasons why this may be true. 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, all programs in my sample implement a BSA at some point in time. The identifying variation therefore comes from *when* programs implement the BSA, instead of *which* programs implement a BSA. Finally, as the performance standard is often uniformly implemented for entire faculties or universities, the performance standards are unlikely to be implemented in response to undesirable trends in specific programs.

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, a more indirect violation of assumption 1 arises when prospective students are discouraged

by the implementation of 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.

To assess the importance of this mechanism, I test for deterrence effects in multiple ways. First, I look at the effects of BSA on the size of programs. I find no effects on average, nor for particular adoption cohorts or specific fields of studies. Next, I zoom in on the municipality level, and ask what happens when formerly popular programs implement a performance standard. Also here, I find that students from the same municipality consistently choose similar programs over time, and this is unaffected by the introduction of performance standards. Finally, I show that there are no significant effects on the characteristics of new students, including their prior grades. Overall, this suggests that - if anything - compositional changes in programs are of limited importance.

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 time 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.[18]

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

Using the 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)). 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.

It is important to note that the estimates do not apply to all programs. The last programs received treatment in 2014.[19] Their treatment effects are not identified, as for these programs no comparable untreated groups are available. Similarly, for the 2012 (2013) cohort, it is not possible to identify the effects of having been treated 2 (1) periods ago, because by that time all programs were treated already. As a result, differences between ATT_l for $l \geq 0$ can be driven by differences in the programs. This seems to be of little importance, however, because in Appendix X I show that the estimates are similar when I only consider programs for whom the effects are identified for all exposure lengths.

5 Results

I present the results in chronological order. First, I examine whether implementing a performance standard deters potential students from applying. Next, I assess the 'bite' of the performance standard. That is, does it affect students' career trajectory after the first year. Following this, I analyze the long-term effects on students' performance in their initial program, as well as their outcomes in higher education and the labor market more broadly. Finally, I investigate heterogeneity.

[19] To be precise, three programs implemented the BSA in 2015 and 4 never implemented the BSA. These are somewhat special programs, and therefore not suitable as control groups for the other programs.

5.1 Deterrence effects

In theory, imposing a performance standard is expected to discourage certain students from enrolling for two reasons. First, students who might prefer less intensive study are now confronted with the decision to raise their effort levels to meet this standard. For some of these students, increasing their efforts may not seem worthwhile, causing them to select another program. Second, as the relationship between effort and grades possesses an element of randomness, the performance standard introduces a risk of dismissal, potentially causing delays in graduation and foregoing earnings. This may also lead certain students to choose an alternative program. In both scenarios, this deterrence effect is expected to be most prominent among students whose expected future performance is low.

I investigate the deterrence effect using several outcomes, with the results summarized in Table 3. All estimates are derived from the difference-in-differences specification detailed in Section 4. The first column displays the mean of the outcome variable in the last pre-treatment year, while the next four columns present placebo estimates for the four pre-treatment periods. The final column shows the Average Treatment Effect on the Treated (ATT), calculated as a weighted average of period-specific ATT estimates for the three post-treatment periods.[20]

The main outcome where an effect may be expected is the average program size. However, the estimates in panel A indicate that imposing a performance standard does not impact the number of students. The estimate itself is small, and the standard error rules out decreases in program size up to 0.08 standard deviations with 95 percent confidence. In Appendix X, I also check heterogeneity in effects by studying effects on particular adoption cohorts or specific fields of study, and find similar effects.

The lack of an effect on program size provides strong evidence against deterrence effects. It implies that if a deterrence effect exists, a comparable number of students must be drawn to the performance standard to maintain a stable program size. In that case, the composition

[20] See Appendix X for the event-study plots that include these period-specific exposure effects.

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 centralized secondary school exam scores. Since grades are available only for the years 2007-2013, I provide pre-trend estimates for up to two years. As shown in panel B, there are no significant effects for any of these variables. [21]

To capture the program fit of new students more accurately, I use a Machine Learning model that combines the variables above with a comprehensive set of additional factors to predict students' graduation chances in specific fields.[22] I then assess whether programs implementing the BSA attract students with higher predicted graduation probabilities in those specific fields. As shown in panel C, the results indicate no evidence that implementing a performance standard increases the quality of new students, even after including course-specific grades in the predictive model.

Finally, in section 6.2, I zoom in on students' program choices at the municipality level. I show visually and through regression results that whenever formerly popular programs 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 all the evidence above, that implementing a performance standard does not deter students from enrolling.

[21] Only the effect on GPA is marginally significant. However, its coefficient corresponds to less than 0.1 standard deviations of the grade distribution and matches that of a placebo estimate, making it difficult to interpret. In Section X, I use synthetic difference-in-differences to compute similar estimates, yielding more precise null effects for this outcome. Additionally, as shown in Panel C of Table 3, there is no effect on program fit, even when grades are included, suggesting at most a limited impact on student quality.

[22] The specifics of the prediction exercise are discussed in section 3.3.

Table 3: Effect of BSA on number and composition of new students

	Mean (SD)	ATT_{-4}	ATT_{-3}	ATT_{-2}	ATT_{-1}	ATT
A. Program size						
Number of students	96.668	-1.422	-3.499	-1.037	-3.875*	-3.803
	(120.953)	(4.672)	(3.616)	(3.055)	(1.977)	(2.641)
B. Student characteristics						
Male	0.481	-0.01	-0.008	-0.005	0.001	0.003
	(0.234)	(0.01)	(0.008)	(0.006)	(0.006)	(0.006)
Age	19.292	-0.011	-0.008	-0.048***	-0.004	0.032
	(0.442)	(0.031)	(0.025)	(0.018)	(0.017)	(0.022)
Previously enrolled in HBO	0.057	-0.002	0.001	-0.004	0.001	0.000
	(0.051)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)
Foreign	0.135	-0.022*	-0.018*	-0.02**	-0.015**	-0.004
	(0.152)	(0.012)	(0.01)	(0.008)	(0.008)	(0.009)
Income percentile father	0.682	0.002	-0.003	-0.001	-0.005*	0.000
	(0.044)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Secondary school GPA	6.471			0.035	0.000	0.032*
	(0.327)			(0.028)	(0.016)	(0.017)
C. Student quality						
Likelihood graduation	0.658	0.002	0.001	0.003	-0.002	0.000
	(0.066)	(0.006)	(0.004)	(0.004)	(0.002)	(0.003)
Likelihood graduation (incl. grades)	0.656			0.005	-0.001	0.000
	(0.068)			(0.005)	(0.003)	(0.004)

5.2 Short run effects of the performance standard

Figure 2 shows that performance standards have large effects on drop-out rates. Since I cannot observe the dismissal status of students, the dropout rate includes both voluntary dropouts and dismissed students. The figure is centered around the pre-treatment dropout rate of 23 percent. The ATT estimate reveals that the performance standard increases dropout rates after the first year by nearly 7 percentage points (25 percent). This indicates that the performance standard leads to the removal of a significant number of students from programs who would otherwise have continued. The total number of students receiving a dismissal is considerably higher than 7%. This is because some students who have not met

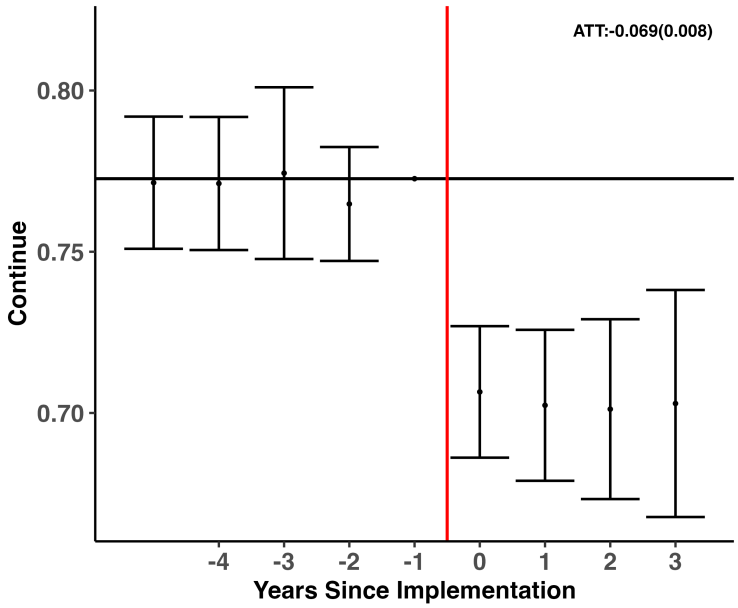


Figure 2: The effect of BSA on dropout after the first year

the standard would also have left without it.[23]

What do students do after they drop out? Table 4 presents estimates for alternative career choices following dropout. Row 1 shows that the rate of switching institutions increases by almost 5 percentage points. Assuming this increase is primarily due to the additional 7 percentage points of dropouts, it suggests that most dismissed students opt to transfer to different institutions.

Next, I explore four mutually exclusive alternatives: (i) switching to a program within the same field, (ii) switching to a program in a different field, (iii) enrolling in a University of Applied Sciences, and (iv) leaving higher education entirely. The final option includes students either exiting education in the Netherlands altogether or entering vocational education. The remaining rows of Table 4 provide these results. Switching to a program in the same field rises by only 1.9 percentage points. This implies that most dismissed students make substantial changes to their specializations. Switching to a university program in a different field shows the largest increase at 2.6 percentage points, while enrollment at Uni-

[23]Ministerie van Onderwijs, Cultuur en Wetenschap (2010) show that about 20% of students did not meet the performance standard between 2005 and 2007.

versities of Applied Sciences and exits from higher education rise by 1.1 and 1.4 percentage points, respectively.

Overall, these results emphasize that implementing a performance standard can cause substantial changes in students' trajectories. Most dismissed students leave their preferred institution and field of study, and some 'downgrade' their level of education or drop out of education.

Table 4: Effect of BSA on education choices after the first year

	Mean (SD)	ATT_{-4}	ATT_{-3}	ATT_{-2}	ATT_{-1}	ATT
Enroll at the same institution	0.85	-0.003	-0.003	-0.001	-0.009**	-0.049***
	(0.066)	(0.006)	(0.006)	(0.004)	(0.004)	(0.007)
Same field (university)	0.053	0.000	-0.001	-0.008	0.001	0.019***
	(0.053)	(0.006)	(0.005)	(0.008)	(0.006)	(0.005)
Other field (university)	0.066	0.000	0.005	0.009**	0.002	0.026***
	(0.042)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
University of Applied Sciences	0.062	0.000	-0.001	0.000	0.004	0.011***
	(0.045)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
Leave higher education	0.046	0.001	-0.002	-0.003	0.001	0.014***
	(0.039)	(0.004)	(0.003)	(0.003)	(0.002)	(0.005)

Notes:

The effects of performance standards are likely not limited to dismissed students. For instance, students who would have persisted regardless might achieve better results with the standard in place. While I do not have access to students' grades or course credits, I do observe long-term outcomes such as graduation rates, time to graduation, and labor market performance, which arguably capture the most important benefits the policy might generate. The analysis of these outcomes is provided in the next subsection.

5.3 Long-term effects of performance standards

The government's goal for the performance standard was to redirect students who are unfit for a particular program toward more suitable alternatives, thereby improving their graduation rates and enrollment durations. This section evaluates whether these goals have been met.

Performance in the initial program. Panel A of Table X illustrates the effects on student performance in the program they initially enrolled in. These estimates are critical from the programs' perspective. The estimate in row 1 indicates that implementing a performance standard reduces the average enrollment duration by approximately 0.2 years (3 months). This reduction is the primary benefit for programs, as they incur costs for each enrollment. However, the performance standard also decreases the graduation rate by nearly three percentage points, suggesting that some students who would have graduated are dismissed. This reduction in graduates imposed a negative consequence for programs since they received payment for each graduate. Whether shorter enrollments but fewer graduates are worthwhile depends on a program's marginal costs and benefits.

The estimates above reflect average effects for all students who started their programs, including dropouts. A key question is whether the performance of the remaining students improved. However, the estimates in rows X and Y show no evidence to support this. Specifically, row Y indicates that performance standards do not affect the average time-to-graduation for those who complete their programs. I also find no effects on the 75th quantile of enrollment duration distribution, suggesting that even among the slowest students, there is no visible improvement. Of course, these estimates may be influenced by selection effects, as performance standards reduce the graduation rate by 3 percentage points (5%), potentially altering the composition of graduates. But if there are selection effects, in the sense that mostly low-ability students are removed from the program, we would expect to see a mechanical decrease in the remaining students' time-to-graduation. The absence of any significant effects therefore suggests that performance standards did not shorten the time-to-graduation for the remaining students.

Higher education performance. Panel B shows that performance standards do not significantly improve students' graduation rates or enrollment durations in higher education. The estimates in rows Y and Z are precise, ruling out improvements greater than 0.x and 0.y standard deviations for these outcomes. Thus, on average, performance standards fail to

achieve their intended goal of reducing enrollment duration or increasing graduation rates. Furthermore, no effects are observed at the 75th percentile of the enrollment duration distribution, suggesting no visible impact even among the slowest students, who were the primary target of the performance standards.

One limitation of the approach taken here is that the estimates capture only the average effects on all students. While this is the most relevant statistic from a policy perspective, it overlooks how dismissal specifically affects the dismissed students themselves. It is possible that dismissed students experience either positive or negative effects, but their contribution to the overall average may be too small to detect. To explore this further, I analyze the effects on students with an ex-ante low probability of graduation — those most likely to be dismissed — in Section 5. Even for this group, I find no long-term effects. This aligns with findings by Vooren et al. (2020), who show with a regression discontinuity design that dismissed economics students from the Netherlands are just as likely to graduate and do not experience study delays.

Labor market outcomes. Another way for the performance standard to generate benefits is to improve labor market outcomes. This can happen when the performance standard redirects students towards more suitable careers or when it improves the skillset of remaining students in their initial programs. Consequently, even when graduation rates and time-to-graduation might remain unchanged, students' labor market prospects could have improved. However, as illustrated in panel C, treated students exhibit similar labor market participation rates and earnings twelve years after their initial enrollment.[24]

In conclusion, while the performance standard significantly influences students' career trajectories, it does not enhance higher education or labor market outcomes. These findings suggest that the performance standard fails to achieve its objective of improving the match quality between students and programs.

[24]As labor market outcomes twelve after enrollment are only available for older cohorts (see Table 1), these results rely on a smaller sample. Nevertheless, the impact of the performance standard on the dropout rate is equally large for these cohorts, suggesting that the treatment is comparable for these periods.

5.4 Heterogeneity

In Table 5, I present effect heterogeneity by program fit and gender. In the first column, I reiterate the primary outcomes for the entire sample. Columns 2 and 3 segment the sample by predicted graduation probabilities. Specifically, I categorize the 50 percent of students with the lowest graduation chances in each program as the 'high risk' group, with the remaining students labeled as the 'low risk' group.[25] Columns 4 and 5 divide the sample based on sex.

As expected, the findings in columns 2 and 3 demonstrate notably larger short-run effects for high-risk students. The effects on drop-out are almost 50 percent higher and their average enrollment duration and graduation chances in their initial program are affected more than that of low-risk students. However, also for the high-risk students, I find no discernible impacts on overall graduation rates, duration, or labor outcomes.

Columns 4 and 5 show remarkable gender heterogeneity. The impact on dropout is more than 60 percent higher for men than for women, and it even exceeds that of high-risk students. The differences in effects among males cannot be fully explained by their capacity to graduate. It thus appears that the performance standards unintentionally targeted men, rather than students unfit for the program. I also find stronger effects on the enrollment duration and graduation rates in the initial program for men. However, I find no impact on general higher education or labor market outcomes for men or women.

The absence of long-term effects may mask significant heterogeneity by field of study or adoption cohort. For instance, being dismissed from a Physics program might have different consequences than being dismissed from a Law program. To explore this possibility, I examine the heterogeneity of effects by adoption cohort and field of study for five main outcomes in Appendix X. The estimates are generally consistent across fields and cohorts, with a few minor exceptions. The only notable exception is the university that first implemented the BSA in 1996, where the effects are generally larger, and both degree attainment and enrollment duration in higher education decrease significantly.

[25] Further details on these groups can be found in Table 2.

Table 5: Heterogeneity by gender or predicted risk group

	Mean (SD)	ATT	High Risk	Low Risk	Men	Women
	(1)	(2)	(3)	(4)	(5)	(6)
Continue	0.773	-0.069***	-0.075***	-0.046***	-0.086***	-0.052***
	(0.092)	(0.008)	(0.012)	(0.011)	(0.009)	(0.01)
Years enrolled in program	3.434	-0.149***	-0.161***	-0.067	-0.194***	-0.101*
	(0.745)	(0.043)	(0.058)	(0.052)	(0.047)	(0.052)
Graduate from first program	0.660	-0.028***	-0.029***	-0.022**	-0.029***	-0.025***
	(0.122)	(0.008)	(0.01)	(0.01)	(0.009)	(0.009)
Years enrolled in HE	6.651	0.006	0.057	0.067	0.011	0.013
	(0.895)	(0.046)	(0.055)	(0.049)	(0.051)	(0.053)
Graduate from any program	0.893	-0.008*	0.002	-0.001	-0.003	-0.010*
	(0.06)	(0.004)	(0.005)	(0.004)	(0.006)	(0.006)

6 Extensions

I extend the analysis in multiple ways. First, I present synthetic difference-in-difference estimates. Next, I test once more for deterrence effects using an alternative approach. Finally, I analyze the potentially adverse effects of performance standards by considering additional survey evidence and medicine prescriptions.

6.1 Synthetic Difference-in-Differences

The results in this paper rely on an unconditional difference-in-differences specification. I check whether the estimates are robust to the Synthetic Difference-in-Differences estimator that automates the process of choosing suitable control groups while maintaining statistical guarantees. The results, reported in Appendix X, are almost identical. This is perhaps unsurprising given that pre-trend violations were at most mild to begin with.

6.2 Testing for deterrence effects: municipality level results

The main results rely on the assumption that changes in program composition have limited importance. To further assess this assumption, I focus on prospective students' program choices at the municipality level and examine to what extent these choices change once previously popular programs implement performance standards.

To fix ideas, consider a simplified example with only two programs: Economics and Mathematics. Between 1994 and 2002, 70% of all students enrolled in Economics. Therefore, in subsequent years, I would expect 70% of prospective students to continue enrolling in Economics. Now, suppose Economics implements a performance standard in 2005, while Mathematics does so in 2008. The key question is: do students switch from Economics to Mathematics between 2005 and 2008? To evaluate this, I predict that before 2005, no students are exposed to a performance standard. After 2005, 70% of students are expected to be exposed to a performance standard, increasing to 100% by 2008. If performance standards deter students, the share of students exposed should be less than 70% between 2005 and 2008. In other words, if students change their enrollment behavior due to performance standards, the actual increase in exposure should be smaller than predicted.

I apply this analysis to all municipalities and programs. Figure 3 shows the predicted and realized fraction of students exposed to a performance standard in the four largest municipalities. This figure helps visualize the identifying variation. Consider, for example, the predicted exposure for students in Municipality 2. Between 2002 and 2009, predicted exposure increases by about 20%, whereas in other municipalities, it rises more quickly. This is because only a few programs popular in Municipality 2 implemented a performance standard during this period. In 2009, however, the predicted exposure jumps to almost 90% because many previously popular programs implemented a performance standard that year. These programs are from a nearby university that adopted the performance standard in 2009.

The realized fraction of treated students closely mirrors the predicted fraction in Figure

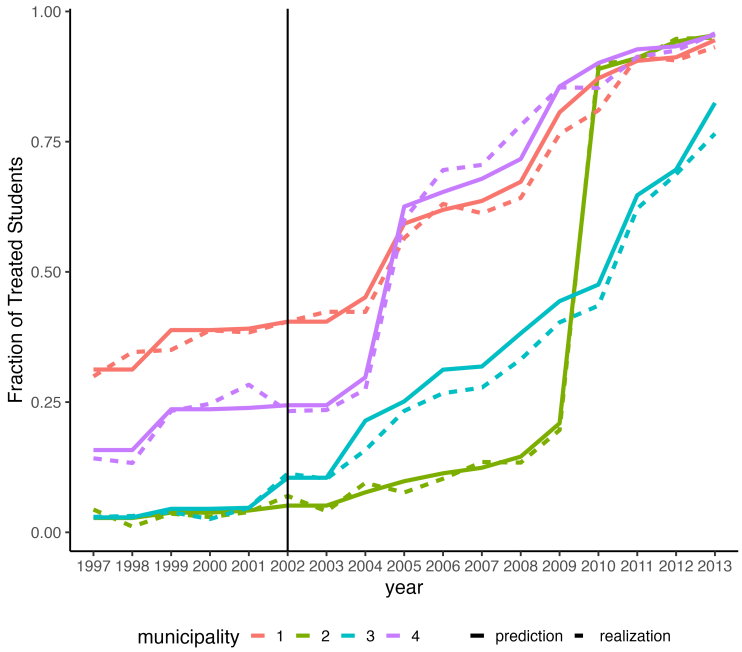


Figure 3: Predicted and Realized Fraction of Treated Students in Four Municipalities

Note: the solid (dashed) lines report the predicted (realized) fraction of students enrolled in programs with a performance standard for four municipalities.

3. Whenever previously popular programs implement a performance standard, the number of students exposed to these standards increases proportionally. In Appendix X, I regress the share of students exposed to a performance standard on the predicted shares across all municipalities. By including municipality and year fixed effects, the analysis ensures that variations in the predicted share of treated students are due to the staggered implementation of performance standards by previously popular programs. The coefficient is remarkably close to 1, indicating an almost perfect alignment between predicted and realized shares across the broader population.

The results imply that if there is a deterrence effect it should (i) attract as many new students as it deters, for otherwise the average enrollment number would change, (ii) it should also attract and deter equally many students *within* municipalities, for otherwise the predicted and realized shares of exposed students would not align, and (iii) the individuals who are attracted or deterred by performance standards do not differ in their gender,

socioeconomic status, or quality.

6.3 Measuring the (dis)utility of performance standards

A performance standard of the type in this paper is an intervention that restricts some students' choices by taking away the option to continue in a program. In the absence of externalities or internalities, reducing choice sets can never increase welfare. The intervention is not free of costs either, because universities face administrative costs and numerous appeal cases. These losses can be justified when performance standards reduce negative externalities. For example, the goal was to reduce the time students spend in education, thereby reducing education subsidies and forgone earnings. However, the main results in this paper indicate that the performance standard does not improve students' time spent in education or labor market outcomes.

Another justification could be an increase in positive internalities. For example, students are known to disproportionally discount the future returns to education, inducing them to invest suboptimal levels of effort in their studies. If that is the case for first-year bachelor students too, then the performance standard can be a welcoming commitment device that gives students a more immediate incentive to increase their effort level.

To test whether the performance standard is perceived as a distortionary intervention or a welcoming commitment device, I surveyed 321 first-year students in two bachelor programs from the same university in the Netherlands.[26] Students are asked five hypothetical choice questions where they trade off monetary gifts and the immediate removal of the performance standard. I use a staircase method that enables me to narrow down the interval around the amount where students are exactly indifferent between the gift and the removal of the performance standard. I also ask for students' probabilistic beliefs about the number of course credits they will have by the end of the year. I use this to measure students' subjective

[26]The sample is not representative of a full class. This research was part of a research participation course, where students could sign up for different experiments. Moreover, by the middle of the year, some students may have dropped out already.

probability of being dismissed. The survey is conducted in the middle of the academic year. At this point, students can have attained at most 30 out of 60 credits already, whereas the performance standard requires them to obtain 48 credits by the end of the year. The exact phrasing of the question and other details of the survey are discussed in Appendix X.

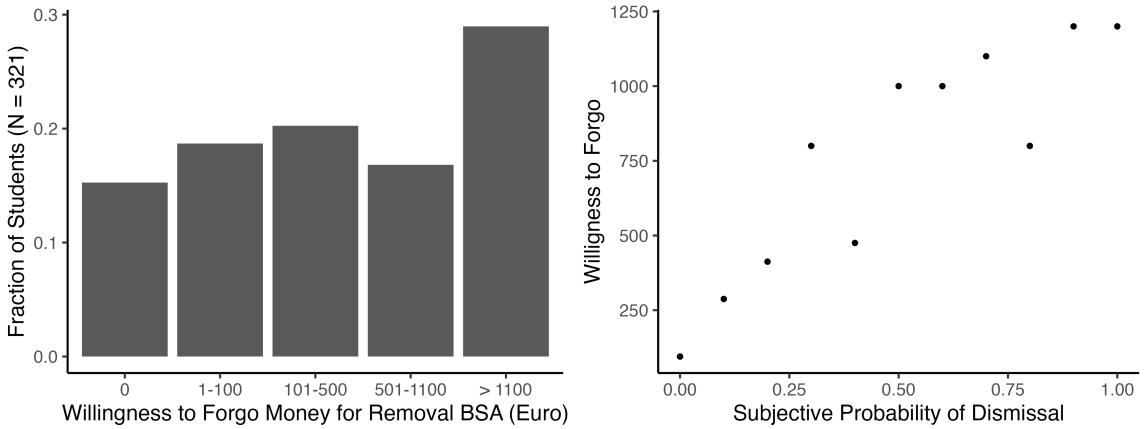
Figure 4 (a) reports a histogram of students' willingness to forgo money for the immediate removal of the performance standard. A remarkable 85 percent is willing to forgo *some* money for the removal of the performance standard. This suggests that very few students perceive the performance standard as a welcoming commitment device because in that case, students would prefer to keep the performance standard instead of a small gift. Moreover, 45 percent of students are willing to forgo over 500 euros, and 30 percent even indicated the maximum amount of 1100 euros. This is more than the average monthly disposable income of students, which is estimated to be 943 euros.

In Figure 4 (b), I plot the median willingness to forgo gifts against students' subjective probability of dismissal. There are two interesting points. First, the median willingness to forgo gifts is positive for all probability bins. This indicates that even students who indicate zero or very low risk of dismissal are often willing to forgo money for the removal of the performance standard. Second, the willingness to forgo gifts increases sharply in the subjective probability of dismissal. This suggests that students who are most likely going to be dismissed do not believe that they are better off in another program.

The findings above indicate that performance standards induce considerable disutility. These results are in line with evidence from Sneyers and De Witte (2017), who use survey evidence from multiple programs and show that implementing a performance standard decreases program satisfaction.

6.4 Medicine usage

A specific way through which the performance standard can harm students is if it unintentionally impacts their mental well-being. This could happen in two distinct ways. Firstly,



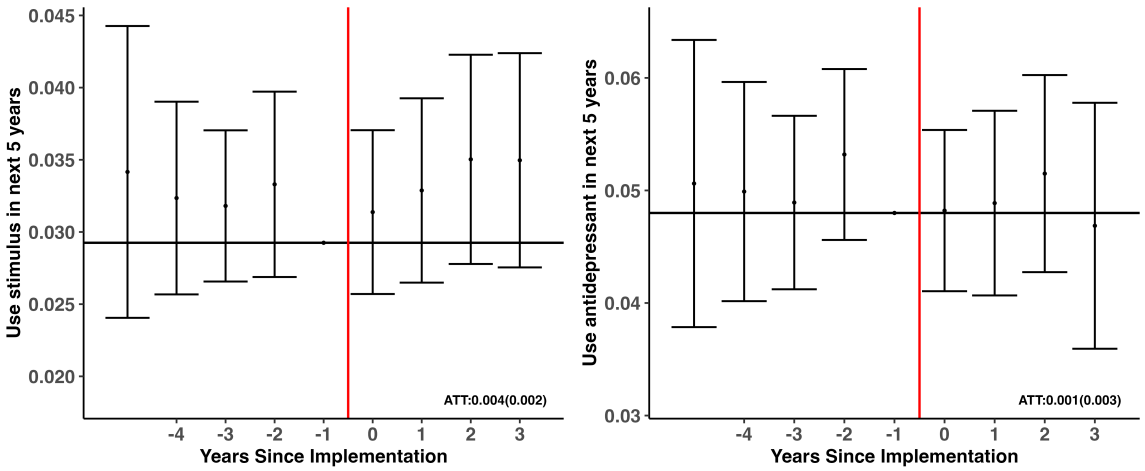
(a) (b)

Figure 4: Willingness to forgo gifts for removal of the performance standard

heightened pressure to perform could lead to increased usage of psycho-stimulants among students. Recent reports highlight a significant surge in Ritalin usage — a psychostimulant often used by individuals with ADHD — among students and associate this trend with the enforcement of binding study advice. Second, the amplified performance pressure may elevate stress levels among students, adversely affecting their overall mental health. Student mental health in the Netherlands has shown a declining trend, and the binding study advice is repeatedly mentioned as a potential contributor (reference).[27] Despite anecdotal evidence and limited-scale surveys addressing student mental health concerns linked to the binding study advice, there exists no definitive causal evidence on its effect.

To fill this gap, I explore the impact of the BSA on two types of medication usage. First, I examine the usage of Ritalin, a psychostimulant commonly used by students to enhance focus. Psychostimulant usage among students has doubled during the period coinciding with BSA implementation in programs. My analysis aims to disentangle whether this association is causal or whether other factors contribute to increased Ritalin usage. Secondly, I investigate antidepressant usage. Acknowledging that mental health impacts might manifest over time, I evaluate psychostimulant or antidepressant usage within five years of enrollment. As the data

[27]In response, the government plans to reduce the performance threshold to a maximum of 30 credits (50 % of all credits) in the first year and another 30 credits in the second year by 2025.



(a) Psychostimulants (b) Antidepressants

Figure 5: Effects of imposing a performance standard on medicine usage within 5 years after enrollment

are available for the years 2006 and later, I restrict the sample to programs that implement the BSA after 2006.

My findings indicate no evidence supporting the notion that the performance standard elevates the usage of psychostimulants or antidepressants. Specifically, the effects in Figure 5 lack statistical significance and rule out effects larger than a 25 % increase in psychostimulant usage and a 20% rise in antidepressant use with 95 % confidence.

While these results are a step towards a better understanding of mental health effects, it is important to acknowledge that my results do not rule out negative effects. Students often buy Ritalin illegally, so actual Ritalin usage could be measured with great error. Moreover, antidepressants are a relatively extreme outcome. The performance standard may affect students' mental health in other ways that do not directly manifest in medicine usage.

7 Conclusion

While performance standards are widely used in education, their effects are unclear. This paper evaluates the implementation of performance standards in the first year of bachelor programs at Dutch universities between 1994 and 2014.

I show that even when a performance standard may substantially affect students' career trajectories, it need not discourage prospective students from enrolling. Moreover, although the government's goal was to redirect dismissed students towards more suitable career paths, students did not benefit in terms of educational attainment or labor market outcomes. Instead, additional survey evidence suggests that performance standards cause considerably disutility. I therefore conclude that academic dismissal policies should be implemented with caution and deserve further scrutiny.

The results also raise new questions. Most importantly, if performance standards of the type considered in this paper do not work, then what does work to improve students' effort and redirect unfit students towards more suitable alternatives? Previous works by Canaan et al. (2022) and Albert and Wozny (2024) suggest that additional support can improve the consequences of academic probation. Similarly, perhaps providing support and career-advice to students who are at risk of dismissal may also alleviate the negative consequences of a dismissal.

References

Albert, Aaron, and Nathan Wozny. 2022. "The Impact of Academic Probation: Do Intensive Interventions Help?" *Journal of Human Resources* 0520–10877R2.

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

Altmejd, Adam, Andrés Barrios-Fernández, Marin Drlje et al. 2021. "O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries*." *The Quarterly Journal of Economics* 136 (3): 1831–1886.

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.

Arpita, Patnaik, Matthew Wiswall, and Basit Zafar. 2021. "College Majors." In *The Routledge Handbook of the Economics of Education.*

Bach, Maximilian, and Mira Fischer. 2020. "Understanding the Response to High-Stakes Incentives in Primary Education."

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 Schnorr. 2022. "Keep Me In, Coach: The Short- and Long-Term Effects of Targeted Academic Coaching.." August.

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.

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.

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.

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

Fletcher, Jason M., and Mansur Tokmouline. 2018. "The Effects of Academic Probation on College Success: Regression Discontinuity Evidence from Four Texas Universities." January.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro. 2021. "Visualization, Identification, and Estimation in the Linear Panel Event-Study Design." August.

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

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

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.

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

ter Meulen, Simon. 2023. "Long-Term Effects of Grade Retention."

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

Ost, Ben, Weixiang Pan, and Douglas Webber. 2018. "The Returns to College Persis-

tence 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.

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.

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.

Wiswall, Matthew, and Basit Zafar. 2021. "Human Capital Investments and Expectations about Career and Family." *Journal of Political Economy* 129 (5): 1361–1424.

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

Zafar, Basit. 2013. "College Major Choice and the Gender Gap." *Journal of Human Resources* 48 (3): 545–595.

Appendix A: Supplementary Results

Table A1: The performance standard threshold in bachelor programs

Credit threshold	Number of programs
30	75
33	1
34	5
35	2
36	53
37.5	34
38	3
39	11
40	38
42	48
45	56
48	25

Notes: This table reports the minimal number of credits that students need to obtain to satisfy the performance standard.

Table A2: Heterogeneity by gender or predicted risk group

	Mean (sd)	ATT -4	ATT -3	ATT -2	ATT -1	ATT
Years enrolled in program	3.434	-0.015	-0.011	-0.045	-0.029	-0.149***
	(0.745)	(0.077)	(0.063)	(0.045)	(0.029)	(0.043)
Graduate from first program	0.66	-0.016	-0.023**	-0.011	-0.007	-0.028***
	(0.122)	(0.01)	(0.009)	(0.007)	(0.005)	(0.008)
Time to graduation first program†	4.097	0.034	0.045	-0.001	-0.001	-0.009
	(0.883)	(0.06)	(0.046)	(0.036)	(0.025)	(0.035)
Years enrolled in HE	6.651	0.048	0.067	0.065*	0.051	0.006
	(0.895)	(0.063)	(0.051)	(0.039)	(0.032)	(0.046)
Graduate from any program	0.893	-0.002	-0.004	0.007**	-0.002	-0.008*
	(0.06)	(0.005)	(0.008)	(0.004)	(0.003)	(0.004)

† : This outcome is conditional on graduating from the first program.