Shadow education in the Netherlands

The position of shadow education in the educational landscape and students' school careers

Jansen, D.

Publication date
2021

3

Chapter 3

Is a more selective exit exam related to shadow education use? An analysis of two cohorts of final-year secondary school students in the Netherlands[iii]

ABSTRACT

The prevalence of private supplementary tutoring (i.e., shadow education) is growing, particularly in nations with selective school exams. The hypothesis that tutoring attendance rises as pressure to perform increases has not yet been tested. Therefore, our research question is: Is students' likelihood to attend shadow education related to the selectivity of the school system? Our study used an opportunity to study this question in the Dutch context, where performance standards on the nationwide secondary education exit exam were raised in 2011. We used data from two cohorts of Dutch exam year students: one before and one after the raised selectivity, hypothesizing that the latter group had a higher likelihood of attending shadow education than the former. Results from Propensity Score Matching (PSM), applied to obtain cohorts as similar as possible on observable characteristics, provide support for our hypothesis. Our results on the Dutch policy change confirm that shadow education use might emerge at key moments of selection. In doing so, our study suggests that policy makers and researchers may want to include shadow education use in discussions on the merits and pitfalls of introducing more stringent performance standards into the education system.

[iii] This chapter is based on: Jansen, D., Elffers, L., Jak, S. & Volman, M. L. L. (submitted). Is a More Selective Exit Exam Related to Shadow Education Use? An analysis of Two Cohorts of Final-Year Secondary School students in the Netherlands.

INTRODUCTION

Shadow education refers to activities that are provided in exchange for a fee and are supplemental to formal schooling (Bray, 2020). Its use in the Netherlands has mirrored international patterns and has grown considerably (Bray, 2020). In fact, annual Dutch household tutoring expenditures increased from 26 million EUR in 1995 to 320 million EUR in 2019 (Statistics Netherlands, 2021). Cross-national studies have related the prevalence of shadow education to the design and functioning of the formal school system (Baker et al., 2001; Zwier et al., 2020). One of the institutional features related to the use of shadow education is selectivity, which refers to the extent to which access to a subsequent level of education is (partly) conditional upon performance on a formal uniform exam (Van de Werfhorst & Mijs, 2010). The proposition is that selectivity puts students and parents under significant pressure to succeed in these exams to ensure their place in higher levels of education. Accordingly, studies on this issue hypothesize that selective educational systems have more students attending shadow education (Baker et al., 2001; Entrich & Lauterbach, 2019; Zwier et al., 2020).

To date, this hypothesis has mainly been tested through cross-national research comparing shadow education use in systems with differing institutional features; for instance, the procedures and timing of the standardized exam (Entrich, 2020; Entrich & Lauterbach, 2019; Zwier et al., 2020). Such analyses have also been applied to regions within countries such as Germany (Guill & Lintorf, 2019) and South Korea (Byun, 2010). These studies consistently report small general correlations between educational selectivity and shadow education use. In these studies, selectivity is usually measured using proxies that are not transposable, or are only weakly transposable, from one country to another (Baker et al., 2001; Zwier et al., 2020). Country-specific studies that focus on selectivity and its relationship to shadow education are often not feasible, as within one country, nearly all students participate in the same testing regime, making valid inferences difficult due to the lack of a control group.

When there is a policy change within a country, it provides a unique opportunity to conduct a study that includes a control group. Such an opportunity arose in the Netherlands in 2011, when the government introduced stricter exam requirements which raised nationwide test standards at the end of secondary education. The Dutch exit exam fits the definition of a “high-stakes” test, as admission to higher education is conditional on student performance on the test (Jackson et al., 2020). Before 2011, previous school grades could be used to compensate for failing the nationwide exit exams (College for Tests and Exams, 2011). After 2011, this possibility was reduced, when the average nationwide exit test scores started requiring a passing mark (i.e., above 5.5 on a scale from 1 [very poor] to 10 [excellent]) with only one five allowed in the core subjects of mathematics, English, and Dutch (College for Tests and Exams, 2011). Thus, the standards to perform on the test changed (Vermeulen et al., 2012, 2013), which increases performance pressure to showcase such ability at one specific "key" moment of selection. Our paper exploits this introduction of a more selective version of the testing regime, and the corresponding increase of performance pressure, to examine their relations to shadow education use. We do so by using data from two cohorts of students: one which completed exams before and the other after the policy change had been introduced. We expect shadow education use to increase in the second cohort (i.e., the one affected by the policy change) compared to the first cohort. By testing this hypothesis, we seek an answer to our research question: Does the likelihood of attending shadow education increase when selectivity, in this case stricter requirements to graduate, in the educational system increases?

BACKGROUND

A brief introduction to the 2011 policy change and shadow education use in the Netherlands

In the Netherlands, children are sorted into hierarchically ordered school tracks at the age of 11 or 12: pre-university education (*VWO,* six years of schooling), senior general secondary education (*HAVO*, five years of schooling), or one of the tracks preparing students for pre-vocational education (*VMBO*, four years of schooling). A *VMBO* certificate provides direct access to senior vocational education, and a *HAVO* or *VWO* certificate provides access to higher education, although a limited number of programs, such as medicine, only admit students that obtained a certain grade point average (GPA) or final exam grade. Dutch higher education is organized according to a two-tier system which separates research universities from universities of applied sciences (i.e., Dutch: *hogescholen*), where the latter is a more practical type of higher education than the one found in research universities. Almost all higher education institutes are publicly funded, with negligible differences in quality between forms of higher education (Sá et al., 2006). Whereas a VWO certificate provides direct access to both university and higher professional education, a HAVO certificate only provides direct access to the latter.

To transfer from secondary to tertiary education, secondary students not only participate in examinations developed by their schools, but also in standardized nationwide exit tests developed by the Dutch Testing and Measurement Institute (CITO) (Schildkamp et al., 2012). To "safeguard the value" of the high school certificate, the Dutch government encouraged schools to reduce discrepancies between the average school exam grade and the exit exam (Vermeulen et al., 2012, p. 1). In light of declining results on the latter (Vermeulen et al., 2012, 2013), the government raised nationwide test standards in 2011. As shown in Figure 5, overall exit exam scores increased between 2010-11 and 2012-13 after the policy change and have since remained relatively stable, possibly related to these grades being calculated based on a norm term that corrects for annual differences in exam length and difficulty (Education Inspectorate, 2013, 2014; Vermeulen et al., 2012, 2013).

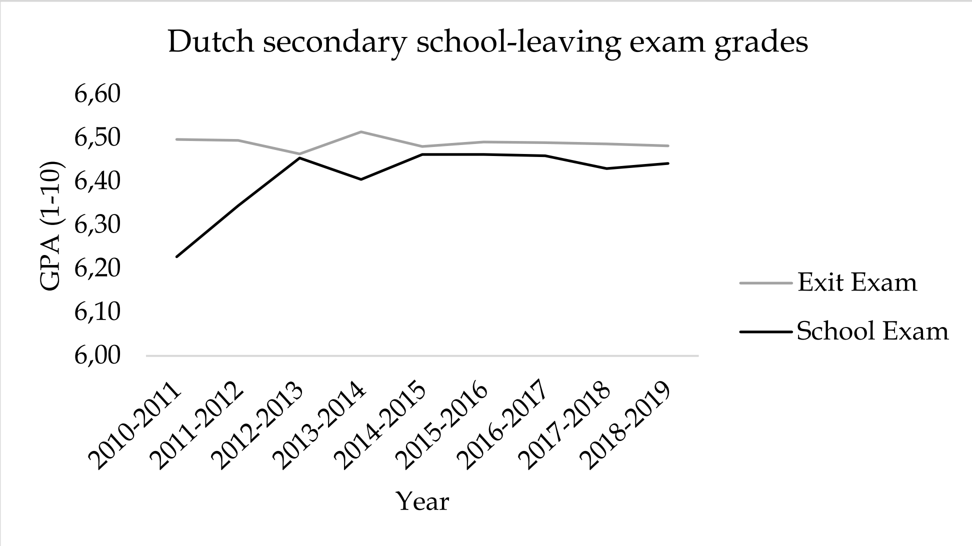

Figure 5. Average results of Dutch secondary school leaving exam (Source: DUO, 2020)

The policy change received significant media attention and raised the question of whether the higher scores could, at least in part, be related to students and parents resorting to private tutoring. Following 2011, some Dutch newspapers reported an increase in shadow education tutoring in the subjects that the policy was aimed at: English, Dutch, and mathematics (Elt, 2012; Maanen, 2014), or in shadow education use in general (Bouma, 2017; Hoogstand, 2012; Kulper, 2017; Vasterman, 2013; Vries, 2013). Individual surveys conducted by tutoring companies show that the policy change was one of the reasons to resort to tutoring for some students. For instance, from the 705 secondary education students (age 12-18) who registered for shadow education, 30% mentioned that the policy change was the main reason for their decision to do so (Dwarkasing, 2012; Lyceo, 2012). Such anecdotal evidence sets the scene for the discussion on, and the formal testing of, the (yet to be confirmed) relationship between a selective institutional context and shadow education use.

Previous work on educational selectivity and shadow education use

Various studies on shadow education have been conducted in countries where selective school exams determine students' opportunity to advance to upper secondary school or university (Byun, 2010; Byun et al., 2018; Hajar & Abenova, 2021; Zhang & Bray, 2017). The exams then function as a "gatekeeper" towards such advancement (Guill & Lintorf, 2019, p. 174). Some studies position shadow education as an institution that is related to neoliberal educational reforms: adaptations of educational system features that emphasize the extension of competitive markets among individuals, families, and institutions (Springer et al., 2016). In East Asia, and increasingly so in Europe and the United States, macro-level neoliberal forces are believed to generate forms of competition on lower levels (i.e., students, parents) (Zhang & Bray, 2017). For instance, when a change to a standardized exam results in students and parents investing efforts towards preparing for such test (e.g., Entrich & Lauterbach, 2019; Park et al., 2016).

Indeed, previous research confirms that increasing demands and social pressures to get admitted to prestigious universities fuels, strengthens, and sustains the prevalence of shadow education. For example, Exley (2020) studied South Korea as an extreme case of how selectivity (in that case the competition for places in middle and high schools perceived to be 'elite' (p. 226)) might relate to shadow education use. South Korea has a long history of examination-related reforms as a result of the growing use of shadow education (Byun, 2010; Lee et al., 2010). With an aim to "reduce the financial burden imposed on families due to shadow education costs" (Lee et al., 2010, p. 100), the South Korean government implemented several measures, including the elimination of the central examinations at upper secondary schools, which some researchers found to have resulted in a significant reduction of shadow education use (Byun, 2010; Kim, 2004).

Researchers have also focused on intra-country differences in existing selective institutional structures and their relation to shadow education, mainly in Germany, Russia, and Japan. First, in Germany, Guill and Lintorf (2019) examined the hypothesis that a correlation exists between standardized tests and the demand for shadow education. By examining shadow education use by final-year primary school students in different German regions which have varying track allocation policies, the authors found an association between high-stakes testing and shadow education use (Guill & Lintorf, 2019). Second, based on their analysis of the Russian educational system, Loyalka and Zakharov (2016) argue that enhanced competition surrounding college admissions fuels the demand for shadow education (see also Jackson et al., 2020; Yastrebov et al., 2018). The Unified State Examination (USE) determines entry to both public colleges and private universities. Because of the importance of this exam, students start preparing for the USE with private tutoring in the years preceding their exam year (Loyalka & Zakharov, 2016). Third, by analyzing data on third- to sixth-grade students in Japan, Matsuoka (2019) found evidence that a high-stakes event, such as an educational transition, fuels college-educated parents to organize their children's time towards preparing for the exam via methods such as shadow education (see also Ireson & Rushforth, 2011).

Building on these studies, we hypothesize that Dutch students taking secondary exit exams under raised standards (i.e., after the policy change in 2011) are more likely to use shadow education to prepare for those exit exams than students taking the exams under the lower standards condition (i.e., before the policy change).

FACTORS INFLUENCING THE USE OF SHADOW EDUCATION

As the testing of the above-mentioned hypothesis requires us to suppress inter-cohort differences, below we detail some of the relevant factors to consider in the selectivity-shadow education relationship.

Previous research has attributed the use of shadow education to achieve better academic results or be allocated to more favorable tracks to a multitude of factors. Particularly during the transition from high school to post-secondary education, it may become a social norm for students from upper-class families to "buy their way into college" (Banks & Smyth, 2015; Smyth, 2009, p. 1). Indeed, shadow education policy-related research often controls for students' socioeconomic status, where students' shadow education use is seen as conditional on their social origin (i.e., relative to their parents' educational attainment; Breen & Goldthorpe, 1997; Byun et al., 2018; Entrich, 2020). According to compensatory advantage models, upper-class parents would be inclined to engage in purposive actions—such as the use of shadow education—to boost their children's school careers (Lee & Shouse, 2011). In other words, irrespective of a selectivity-based policy, high-socioeconomic (SES) students are more likely to attend shadow education, mainly when they attend high-SES schools (Matsuoka, 2015). This is especially the case for students with low performance, which could compromise their chances of attending higher education. Indeed, earlier research confirms the importance of controlling for both SES and performance in shadow education related research (Guerrero, 2020; Huang, 2020). Some researchers have also found that shadow education use is more likely among boys from low-SES families than their female counterparts (see Entrich & Lauterbach, 2019). Also studying predictors of shadow education use, Entrich (2020) argues that, in comparison to non-academic tracks, academic tracks expose students to an environment with increased cognitive demands, making shadow education use more likely in academic tracks. Lastly, shadow education use might also be influenced by one's motivation, such as one's desire to succeed or overall school performance (Guill et al., 2020b).

PRESENT STUDY

Previous research proposes a relationship between the selectivity of educational systems and shadow education use. So far, this relationship has been tested through cross-country comparisons, with proxies that are often not transferable from one country to the other. The lack of a control group limits the possibility to advance knowledge on country-specific effects. In the Dutch institutional context, identification of the relationship between selectivity and shadow education is aided by a nationwide policy change in 2011 that increased the need for good results in accessing higher education. As all schools and regions were required to implement the policy, this change provides a clear cut-off between the two conditions, which we interpret as higher versus lower performance pressure. We hypothesize that, after suppressing potential inter-cohort differences, students in the post-policy cohort will have a higher likelihood of using shadow education than those in the pre-policy cohort. If the policy change results in parents resorting to shadow education to improve their children's scores on these tests (Byun, 2010; Guill & Lintorf, 2019; Kim & Chang, 2010; Lee et al., 2010), this may lead to an increase in inequality in performance with less favorable outcomes for students whose parents do not have the means to resort to shadow education (Choi & Park, 2016).

METHOD

Data

A subsample of a national cohort study (i.e., COOL$^{5\text{-}18}$ study; Hulshof & Timmermans, 2016; Hulshof et al., 2015) was used in this paper, namely two waves among students in the final year of senior, general secondary education (*HAVO 5*), and pre-university education (*VWO 6*). The researchers of the COOL$^{5\text{-}18}$ study applied two-stage, hierarchical sampling, where schools were sampled first, then students within schools were sampled (Hulshof & Timmermans, 2016; Hulshof et al., 2015). The first wave, in 2010-2011, included 4,530 students, and the second wave, in 2013-2014, included 10,332 students. The large difference in sample size could be partially explained by the second wave featuring more schools willing to participate (i.e., 75 and 113 schools in the first and second waves, respectively), but also more schools in the second wave where entire classes - rather than one or two students - participated in the study. A selection of students from both waves (i.e., the so-called *target students*) also participated in a previous wave of data collection three years before their exam year, where questions were asked about students' socioeconomic background or their school motivation. These target students are of particular interest here, because using ninth-grade data allows us to be certain that the covariates used in the analysis are measured prior to the policy change, rather than during or after, as covariates that are measured after treatment assignment cannot act as confounders of the treatment allocation process. After limiting the data to only the target students, 6,878 students remained. See Keuning et al. (2012a), Keuning et al. (2012b), Keuning et al. (2015), and Keizer-Mittelhaëuser et al. (2015) for more information on the aims and execution of the COOL$^{5\text{-}18}$ project.

VARIABLES

Dependent variable

Shadow education use was the dependent variable, measured by asking students if they took private tutoring or not to prepare for their upcoming exams. This question was a binary variable (yes = 1, no = 0), so we could not differentiate between tutoring for math, English, Dutch, or other subjects. The measurement of shadow education did not change across the cohorts.

Independent variable

A dummy variable was added to each database to indicate the cohort (0 = pre-policy, 1 = post-policy).

Controls

To apply PSM, the untestable assumption of ignorability must hold: all potential confounding covariates must be included in the propensity score model. In this study, as in most education-based studies (see Fan & Nowell, 2011), including all covariates is difficult due to the large number of covariates theoretically and empirically linked to shadow education use. Our selection of covariates is based on methodologically similar studies (e.g., Byun, 2010), leading us to include student-level covariates (i.e., SES, baseline performance before engaging in shadow education use, ability, motivation, exam track, subject profile, and gender), the mean baseline performance of the class, and school SES.

SES was measured through parental educational level when students were in the 9th grade. Mothers and fathers answered separate questions regarding their highest completed education and earned certificate (Hulshof & Timmermans, 2016; Hulshof et al., 2015). The low category included parents with no diploma or a primary school diploma. The average category included parents with secondary or vocational education, and the high category included parents with higher professional education or university diplomas. The variable SES was constructed by taking the highest education completed by either the mother or the father. The SES variable was aggregated at the school-level as well. Gender was a binary variable for boys and girls, as was exam track for pre-university (VWO 6) and senior general secondary education students (HAVO 5). Students' subject profile was measured using a binary variable, which was one if the student followed a relatively science-related track and zero otherwise. For performance, standardized tests of language and mathematics were administered focusing on spelling and arithmetic, respectively, with scores ranging from 0 to 100. We aggregated this performance to the class level, although this performance is also investigated at the student level. For intellectual ability, the Dutch intelligence test for educational purposes (see Keuning et al., 2012a, 2012b) was used, which includes six tests focused on verbal and symbolic problems. For student motivation, a Dutch version of the Inventory of School Motivation (ISM) was used (McInerney & Ali, 2006). Consistent with Byun's (2010) focus on intrinsic motivation, we only examined students' performance and mastery-related motivation, measured on a 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree). Performance motivation was measured by items such as, "I work harder if I'm trying to be better than others," and mastery motivation related to items such as, "I like to see that I am improving in my schoolwork" (see, e.g., McInerney & Ali, 2006 for all items). All numerical variables were grand mean centered before analyses.

PROPENSITY SCORE MATCHING

To compare the two cohorts, we applied PSM, which enabled the testing of our hypothesis. PSM is a technique to reduce a set of confounding variables into a single score ranging from 0 to 1 (Austin, 2008; Ho et al., 2007; Luellen et al., 2005; Rosenbaum & Rubin, 1983). In this study, the propensity score is the conditional probability of a student being in the post-policy cohort. Below, we describe how we handled missing data and the steps we took to implement the method.

Missing data

From the 6,878 students in the dataset, 4,376 contained missing data on one or more of the variables of interest. Missing data patterns explored with the finalfit package in R (Harrison et al., 2020) revealed that the missing data of SES, subject profile, intellectual ability, motivation, and gender were unrelated to the use of shadow education, meaning for these variables the data was missing at random (MAR). For the missingness of performance, however, a relation to shadow education use was present (p < .01), meaning for this particular variable, the data was not missing at random. When engaging in multiple imputation across five, ten or twenty datasets, and averaging the propensity score across these datasets (Mitra & Reiter, 2016), results were not substantially different from complete-case analyses. As argued by White and Carlin (2010), in cases of MAR, multiple imputation does not always offer statistical advantages over complete cases. Thus, we proceeded with the analyses of the students for whom we have complete data (n = 2,502), but report supplementary results for the whole sample of students as well, obtained using the Mice (Van Buuren & Groothuis-Oudshoorn, 2011) and MatchThem (Pishgar et al., 2020) package in R.

Analyses

Analyses were conducted in five steps: (1) set up a propensity score model, (2) conduct matching and assess balance across cohorts, (3) check clustering of shadow education use at the class and school level, (4) fit a multilevel logistic regression model to estimate the outcome, and (5) conduct a robustness check. The appendix contains a link to the data and R script.

First, we estimated a propensity score for each individual based on a single-level logistic regression model where all student-level covariates predict the cohort variable using the MatchIt (Ho et al., 2007) package in R statistics software (R Core Team, 2021). Second, students in the post- and pre-policy cohort were matched based on their propensity score. There are several ways to conduct this matching, where the most used method is greedy nearest neighbor matching (i.e., each student is matched to the best option from a pool of matches) (Rosenbaum & Rubin, 1983). Unmatched control cases are then excluded from the analyses, which is why nearest neighbor matching can result in relatively poor matches (Austin, 2008; Ho et al., 2007). Full matching, on the other hand, takes one student from the post-policy cohort and matches him or her with more than one student in the pre-policy cohort (Hansen, 2004; Ho et al., 2007; Sekhon, 2008). Full matching does not refer to matched pairs, but rather matched sets, where each student is assigned a weight that can be used in further analysis (Hansen, 2004). Full matching is optimal in that sense; it minimizes the overall differences between matched sets (see Stuart, 2010 for an overview of matching methods). To retain sample size, we conducted our matching using full matching. We used the nearest neighbor matching as a robustness check (Fan & Nowell, 2011; Ho et al., 2007). We assessed the obtained balance both visually, by plotting the distribution of the propensity scores, and numerically through standardized mean differences. A standardized mean difference below 0.10 between the treated and untreated groups is considered small (Stuart, 2010).

After balance was achieved, our third step was to check the clustering of shadow education use at the class and school level. In doing so, we account for clustering in the second stage of the analysis (Li et al., 2013). For this, we fit three two-level intercept-only logistic regression models: one with only a random intercept at the class level, one with only a random intercept for the school level, and a three-level random intercept model at both the class and school levels. After fitting the models, we selected the best-fitting model based on two criteria: (a) the Akaike Information Criterion (AIC), and (b) Bayesian Information Criterion (BIC). The model with the lowest AIC and BIC was considered as the best-fitting model. Fourth, using the best-fitting model, we fit a multilevel logistic regression model with the propensity score-based weights supplied to the lme4 package (Bates et al., 2011). We fit a multilevel logistic regression model with only the cohort variable, a model with the cohort variable and student-level covariates, and a model with student-level and higher-level covariates. Significance is tested at a = .05, or if the 95% confidence interval of the odds ratio (OR) does not include 1.

RESULTS

We present our results in the five steps that align with the order of our analyses. First, we present the sample before and after matching. Second, we assess the obtained balance after matching. Third, we check the dependency of shadow education use across classes and schools. Fourth, we estimate the relationship between the policy change and shadow education use. A robustness check is presented as a fifth and final step.

Descriptive statistics before and after matching

From the 2,502 students in the sample, 808 were in the pre-policy cohort and 1,694 were in the post-policy cohort. 29.33% and 32,17% of students indicated that they used shadow education in the pre- and post-policy cohort, respectively. Thus, based on these raw numbers, the use of shadow education seems to be higher in the post-policy cohort. Of the students who use shadow education, the majority (n = 1375) have high-SES background, but the percentage of high-SES students resorting to shadow education did not increase across cohorts (60.76% and 50.82% in the pre- and post-policy cohort, respectively). Table 4 shows other descriptive statistics before and after matching. As shown, before matching, there are considerable differences between the post- and pre-policy cohort in terms of motivation, in favor of the post-policy cohort (d = .16), which become smaller (d = .02) after matching. Moreover, the pre-policy cohort has lower performance than the post-policy cohort (d = -.12), a difference that is also suppressed after matching (d = -.07).

Assessing obtained balance after matching

After matching, all standardized mean differences fall below the threshold of .10, therefore, we can be confident balance has been achieved. For some of the covariates (e.g., performance), the percent balance between improvement was negative, meaning the balance between the pre-and post-policy cohort worsened, requiring further inspection of the balance. Next to the standardized mean differences before and after matching, which are shown in Table 4, we also screened variance ratios as indicators of remaining differences in the matched samples, independent of the sample size (Ho et al., 2007). A variance ratio of 1 indicates good matching, whereas values below or above 2 are indicative of a balanced or unbalance sample, respectively. The variance ratios ranged from .89 (performance) to 1.18 (intellectual ability), providing support that the matching procedure was successful. Matching the students using nearest neighbor matching did not result in better matches in terms of standardized mean differences and variance ratios.

Table 4. *Standardized mean differences before and after matching*

	Before Matching			After matching		
	M-C	*M-T*	*d*	*M-C*	*M-T*	*d*
SES: low	.10	.12	.02	.11	.12	.01
SES: average	.12	.10	.06	.12	.11	.02
SES: high	.35	.31	.09	.35	.33	.06
Performance	.53	.59	-.12	.53	.56	-.07
Profile: non-technical	-.16	-.09	-.01	-.16	-.96	.08
Profile: technical	.75	.69	.14	.75	.79	-.09
Intellectual ability	.25	.31	-.14	.25	.21	.09
Motivation	-.15	1.88	-.16	-.15	-.46	.02
Exam track: havo	-.00	.01	-.03	-.00	-.00	.00
Exam track: vwo	.67	.45	.45	.67	.67	-.02
Gender: boy	.33	.55	-.45	.33	.33	.02
Gender: girl	.44	.44	-.00	.44	.43	.01

Note. M = Mean, T = treatment (n = 1,694), C = control (n = 808).

Checking the dependency of shadow education use across classes and schools

As shadow education use may vary across classes and schools, failure to account for such dependency might lead to omitted variable bias. Table 5 shows the fit indices for the three empty models (i.e., without predictors) we compared. The model with nested random effects outperforms the model with only a random intercept at the class level ($\chi2(1)$: 28.37, $p < .05$). The intraclass correlations (ICC) of the nested model show that .08 and .07 of the variances in shadow education use is at the class and school level, respectively.

Table 5. *Fit statistics for models (fit rank in parentheses)*

Model	AIC	BIC	χ(2)	df	*p*
Empty model (class)	2685.69 (2)	2697.34 (2)			
Empty model (school)	2692.85 (3)	2704.5 (3)	.00		
Empty nested model (class and school)	2666.48 (1)	2683.96 (1)	28.37	1	0.0000

Note. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. Lower values indicate better fit. df = degrees of freedom

Relationship between policy change and shadow education use

Table 6 shows the results from the logistic regression models with a nested random intercept for class and schools. It is important here to only interpret the estimate of the cohort-shadow education relationship because the covariates have already been adjusted for through matching. As shown, being in the post-policy cohort relates to the use of shadow education (β = .45, p < .05, OR = 1.57 ([95% CI: 1.07, 2.30]), suggesting that the likelihood of attending shadow education is 1.57 times higher for students that took their exam after 2011. The results also show that the estimates remain stable when including student- or higher-level covariates, which indicates that the matching was successful (Ho et al., 2007). The full model (Model 3) explains 13% of the variance in shadow education use.

Table 6. *Parameter estimates, standard errors, and fit statistics of the multilevel (random intercept) logistic regression models predicting likelihood of attending shadow education*

	Model 1		Model 2		Model 3	
	B	*SE B*	B	*SE B*	B	*SE B*
Constant	-1.23***	.18	-2.00***	.26	-2.05***	.26
Post-policy cohort	0.45**	.20	0.37***	.00	0.43***	.00
Student-level covariates	No		Yes		Yes	
Higher-level covariates	No		No		Yes	
ICC (class)	.08		.08		.08	
ICC (school)	.07		.06		.05	
N	2,502		2,502		2,502	
Akaike Inf. Crit.	2,662.88		2,588.37		2,588.37	
Bayesian Inf. Crit.	2,686.18		2,658.27		2,658.27	

Note. ICC = intraclass correlation. Significance level = ** *p<0.05; *** p<0.01*

Supplementary results

When using multiple imputation (n = 6,878) instead of using complete cases to handle the missing data, the relationship between the post-policy cohort and shadow education use remains significant, when pooled across twenty datasets (β = .42, p < .05, OR = 1.53, [95% CI: 1.22, 1.91], $n_{pre\text{-}policy\ \mathrm{cohort}}$ = 1,706, $n_{post\text{-}policy\ \mathrm{cohort}}$ = 5,172). As these estimates are (nearly) similar to those obtained using complete cases, we can be confident that the found estimates are robust to the way missing data was handled.

DISCUSSION

We set out to test the hypothesis that students with stricter requirements to graduate (i.e., higher selectivity) will have a higher likelihood of using shadow education compared to their counterparts in less selective conditions. Our data provides support for this hypothesis in a matched sample of Dutch final-year secondary school students. The 2011 Dutch policy change can be interpreted as a generic case of introducing greater selectivity into the educational system, which, based on our analysis, appears to relate to students and parents investing more effort towards preparing for the exam by using shadow education. This assertion had thus far—in the Netherlands, but also in other countries where shadow education is not yet widespread—remained unconfirmed due to the lack of a baseline against which a change in selectivity can be studied.

Whereas PSM enables us to identify a relationship between selectivity and shadow education use for a matched sample of exam-year students in the Netherlands, this approach does not allow for causal claims about the relationship between selectivity and shadow education use in general. In other words, PSM is useful to reduce selection bias, but it does not function as a magic solution towards eliminating all of it (King & Nielsen, 2019). Thus, the relationship between selectivity and shadow education use remains uncertain, particularly as shadow education use among Dutch secondary education students continued to increase after 2012 (Statistics Netherlands, 2021). The anecdotal evidence we point to the introduction does, however, indicate that providers of shadow education report a peak in subscriptions in 2011-2012. Nonetheless, other factors, such as teacher and parental expectations, the use of shadow education by peers (Smyth, 2009), or school policies that encourage the use of shadow education by offering providers places within the school may also contribute to a growing use of shadow education. Future studies may extend our approach by allowing respondents to reflect on their motives for using shadow education, to shed further light on the weight that exam-year students and their parents put on exam requirements relative to other factors when resorting to shadow education.

When looking at our findings in relation to previous studies, our findings corroborate that of Guill and Lintorf (2019) for the German case, but differ in some respects as, unlike Germany, where transfers from vocational to higher education are increasingly being allowed, pathways to university are rather limited in the Netherlands. Moreover, we studied a sample of upper secondary education students, whereas previous researchers either focused on final-year primary school students or those in middle school (Byun, 2010). Our findings for a sample close to graduation substantiate what Stevenson and Baker (1992) once asserted: the existence of shadow education is "tightly coupled to the organization of transitions both within schooling and from school to the workplace" (p. 1655). Since 1992, when Stevenson and Baker published their study, the world has witnessed a continuing educational expansion where students and parents increasingly strive to move up the ranks in the educational "arms" race (Halliday, 2016). In this regard, a classic prisoner's dilemma may follow, where parents are aware of the pressures on their children, and many would like to avoid shadow education, but they feel they cannot do so due to the intensified competitive arena induced by certain education policies (Exley, 2020, 2021; Zhang & Bray, 2017).

From a broader perspective, as many education policy initiatives have aimed to lessen, or perhaps control, the use of shadow education (see Edwards et al., 2020; Lee et al., 2010; Piao & Hwang, 2021; Zhang & Bray, 2017), our study presents the case of increased shadow education use resulting from an adaptation of an existing instrument: a national exam. As this was a governmental education initiative in which shadow education was absent from the policy agenda, our study shows that shadow education use can be a 'by-product' of a policy change related to an educational system's selective structure. Thus, the assertion that standardized tests reduce the influence of parental background on students' school careers (Van de Werfhorst & Mijs, 2010) may warrant reconsideration in terms of the recent growth of shadow education. Our findings speak to the relevance of including shadow education in discussions on the merits and pitfalls of standardized testing and performance pressure. Our study underlines the value of a broader scope in educational research than one that is limited to regular schooling. A broader scope can help us understand how the various components of the educational landscape (and the changes therein) shape students' educational trajectories.