DOI: 10.1111/cdev.13985

CHILD DEVELOPMENT

EMPIRICAL ARTICLE

Distressing testing: A propensity score analysis of high-stakes exam failure and mental health

Kathryn Christine Beck[1,2] | **Helene Lie Røhr**[3] | **Bjørn-Atle Reme**[1] | **Martin Flatø**[1]

[1]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway

[2]Department of Sociology and Human Geography, University of Oslo, Oslo, Norway

[3]Kristiania University College, Oslo, Norway

Correspondence
Kathryn Christine Beck, Centre for Fertility and Health, Norwegian Institute of Public Health, Postbox, 222 Skøyen, N-0213 Oslo, Norway.
Email: kathrynchristine.beck@fhi.no

Funding information
Norges Forskningsråd, Grant/Award Number: 262700 and 314562

Abstract

This study used rich individual-level registry data covering the entire Norwegian population to identify students aged 17–21 who either failed a high-stakes exit exam or who received the lowest passing grade from 2006 to 2018. Propensity score matching on high-quality observed characteristics was utilized to allow meaningful comparisons ($N = 18{,}052$, 64% boys). Results showed a 21% increase in odds of receiving a psychological diagnosis among students who failed the exam. Adolescents were at 57% reduced odds of graduating and 44% reduction in odds of enrolling in tertiary education 5 years following the exam. Results suggest that failing a high-stakes exam is associated with mental health issues and therefore may impact adolescents more broadly than captured in educational outcomes.

Research into educational inequalities has pointed to widening relative disparities in health across educational groups (Goldman & Smith, 2011; Lorant et al., 2018; Mackenbach et al., 2015; Strand et al., 2014). In an effort to understand the widening educational inequalities in health, Mirowsky and Ross (2008) established the Rising Importance Hypothesis which suggests that education erodes health at a faster rate for younger cohorts than for older cohorts (Lauderdale, 2001; Mirowsky & Ross, 2008). Resulting from a shift to more knowledge-based societies, education may play a larger role in the attainment of health protective resources and thus the relation between education and health may strengthen across cohorts. In this transition, students may experience larger educational demands and school-related pressure. The Educational Stressors Hypothesis suggests that the increasing psychological burden on adolescents is in part a result of rising school-related stress (Högberg et al., 2020; West & Sweeting, 2003). As educational stress rises through the use of testing and evaluation practices, performing poorly on assessments and therefore failing to complete upper secondary school may place a significant burden on students' psychological well-being (Högberg et al., 2020; West & Sweeting, 2003).

High-stakes testing may contribute to an increase in school-related stress, as these exams have far-reaching consequences for the students' educational trajectory (Högberg & Horn, 2022). Examples of high-stakes exams include those which are used to determine placement in educational tracks, grade retention, entry into certain study programs, and graduation from and access to higher educational levels. In the case of high-stakes exit exams, there are arguments to be made for the benefits of increasing the accountability of educators for students' performance, and the incentivizing of students' continued participation in learning (Högberg et al., 2021; Papay et al., 2010). It is also argued that these high-stakes exams provide an objective opportunity for students to

Abbreviations: GPA, grade point average; ICPC-2, International Classification of Primary Care version 2; IRR, incidence rate ratio; NEET, not in employment, education, or training; OR, odds ratio; PSM, propensity score matching; SMD, standardized mean difference.

Bjørn-Atle Reme and Martin Flatø should be considered joint senior author.

test their cumulative learning through a standardized exam, without the potential bias of grades given by their teachers. Opponents of these exams, however, argue that such exams place undue stress on students, particularly on marginal students already struggling in the education system. These exams may also perpetuate social stratification by placing the largest burden on students from low socioeconomic backgrounds and allocating students into educational and labor market social categories by constraining social mobility (Högberg & Horn, 2022).

Some literature has examined the possibility of cultural bias in standardized tests, where majority students are inherently advantaged in taking such exams. These exams may be constructed in such a way that minority students may respond to the exam differently than the majority population. Additionally, the interpretation of a student's submission may reflect differences in cultural background between students and assessors, resulting in an invalid measurement of student performance (Helms, 2004). Students taking the Norwegian written exam have chosen either Nynorsk or Bokmål as their main language and write the exam in the Norwegian language of their choice, to provide equal opportunities. Instructions are given in both language forms. National minorities with a Sámi language, Kven, or Finnish as first language may be exempted from the exam. Furthermore, high-stakes exit exams may contribute to widening inequalities in health, as failing such exams may act as a "double disadvantage" where these adolescents are at both a lower likelihood of graduating from upper secondary school and increased likelihood of mental health disorders.

We may conceptualize the connection between high-stakes exam failure and mental health through the theory of learned helplessness. Developed by Seligman and colleagues in the late 1960s and early 1970s, this theory posits that when exposed to negative outcomes or stressors, individuals may believe that their behavior will not make a difference and they do not have control over their life outcomes (Seligman, 1972). Three deficits are hypothesized to result from learned helplessness: these are motivational, cognitive, and emotional. Motivational deficit refers to the lack of a response to reverse or rectify a negative situation even when such opportunities are available, while cognitive deficit refers to the idea that the circumstances of one's life are beyond their control. Lastly, emotional deficit is the resulting depressed state from such negative situations perceived to be beyond one's control. This theory was expanded by Abramson et al. (1978) using attributional theory, which hypothesized that individuals react differently to negative events, and the way in which individuals attribute or interpret this event will impact their likelihood of experiencing learned helplessness and subsequent depression. They proposed that individuals classify negative events on three scales; from internal to external: whether the cause of the negative situation is related to personal factors or situational factors, stable to unstable: whether the cause is stable across time or not, and global to specific: whether the cause is consistent across contexts or unique to the specific situation (Abramson et al., 1978).

This model of learned helplessness is theorized to be especially linked to depression, as it theorizes that when a highly desired outcome, such as passing an exam, is believed to be improbable, an individual will believe that their actions cannot have any significant impact on the outcome, leading to depression (Forgeard et al., 2011). Therefore, we may expect that students who failed a high-stakes exam may experience a feeling of loss of control, as the failure of the exam may signal to students that they are not in control; even if they tried to pass the exam, they were unable to change the outcome. This sensation of loss of control may then result in depression and further loss of motivation and effort, leading to negative effects on both mental health and future academic outcomes. Thus, high-stakes exam failure may impact mental health and academic outcomes through many mechanisms related to learned helplessness. Individuals may feel that although they tried to pass the exam, they were unable to obtain a passing grade. If these individuals attribute this failure to internal, stable, or global causes, they may be at a higher risk of depression and be less likely to retake the exam, finish upper secondary school, and enroll in further education. For example, if individuals attribute the exam failure to personal factors, such as intelligence, they may be less likely to invest the additional resources needed to retake and pass the exam, leading to lower educational attainment, as well as poorer mental health.

Prior research on the consequences of high-stakes exams, however, has primarily focused on academic or employment outcomes (Anderson, 2022; Andresen & Løkken, 2020; Caves & Balestra, 2018; Machin et al., 2020; Ou, 2010; Papay et al., 2010). In the United States and the United Kingdom, studies have found causal effects of failing high-stakes exams on enrollment in further education and graduation and dropout rates in upper secondary school (Machin et al., 2020; Ou, 2010; Papay et al., 2010). In studies by Ou (2010) and Papay et al. (2010), researchers found that low-income and minority students are especially at risk for dropout following failure on high-stakes exit exams in the United States, suggesting that these exams may increase educational inequalities.

Fewer studies have examined mental health related outcomes (Högberg & Horn, 2022; Kumandaş & Kutlu, 2010; Wang, 2016). Högberg and Horn (2022) found an increase in school-related stress due to high-stakes testing, and Kumandaş and Kutlu (2010) found higher levels of test anxiety and fear of underachievement in preparation for a high-stakes exam. Finally, Wang (2016) found that students who performed worse than expected on the college entrance exam in South Korea experienced an increased risk of suicidal ideation. This suggests that high-stakes

exams may not only impact the mental health of students who fail, but also among high-achieving students scoring lower than expected.

Understanding the broader impacts of failing these high-stakes exams is important for policymakers looking to reduce high school dropout and improve adolescent mental health. In Norway, 13% of students in academic tracks fail a written exit exam during their final year of upper secondary school. This proportion ranges from 10% to 16% by year, reflecting fluctuations in the difficulty of the exams. These exams are given in May and June, after applications to tertiary education programs are submitted and right before graduation from upper secondary school. Therefore, failing an exam in the final year of upper secondary school often results in a delay in qualifying for tertiary education of at least 1 year. Despite the non-negligible proportion of students who fail these exams, little is known about their impacts beyond academic outcomes. Further research is therefore needed to establish whether there is an association between high-stakes testing and mental health in the Nordic context, and whether these exams may impact students more broadly than through short-term test anxiety and stress. This study therefore contributes to the existing literature examining the association between failing a high-stakes exam and adolescents' mental health, graduation from upper secondary education, and enrollment in tertiary education in the Norwegian setting.

Norwegian school and exam system

The education system in Norway is a publicly funded service provided for all children. Children attend primary and lower secondary school (referred to as compulsory education) from the calendar year in which they turn six until graduating during the year they turn 16. Following completion of compulsory schooling, students may either leave education or enroll in upper secondary education, and very few choose to end their education following compulsory school. Students take a final exit exam at the end of compulsory schooling; however, students cannot fail this exam, and the consequences do not extend beyond the implications of the exam grade on the students' lower secondary school grade point average (GPA).

When applying for admission into upper secondary education, students have a choice between multiple vocational and academic study tracks. Academic tracks last 3 years and qualify the student for tertiary education upon completion, while vocational tracks usually last 4 years, with 2 years of schoolwork and 2 years of apprenticeship resulting in documented competency in the selected occupation. About 50% of the students enroll in a vocational education track after lower secondary school. For this study, we focus on students enrolled in academic tracks only.

In upper secondary school, the students must take various exams each year including oral/practical exams which are locally given, and written exams which are centrally given, both of which are held in late May through early June. The locally given exams are administered at the county level, whereas the written exams are administrated centrally by the Norwegian Directorate for Education and Training (UDIR). These exams are comprehensive and the same exams are given nationwide on the same day, and thereafter, they are anonymized and graded by a randomly assigned external evaluator. Exams and teacher-assessed course grades are evaluated on a discrete scale from 1 to 6, with 1 being a failing grade, 2 as the lowest passing grade, and 6 as the highest grade. Throughout upper secondary school, students are randomly selected to take exams, whether oral, practical, or written, in several courses where the teacher-assessed course grade may be on the students' diploma (called a graduating course). One exception is the Norwegian or Sámi first language written exam, which all students on the academic tracks are required to take and pass in their final year of upper secondary school before graduating.

If a student fails or is unable to take an exam, they are allowed to progress onto the next year and are given the opportunity to retake the exam; however, during the third (final) year, these exams are more high stakes. This is because students are required to take and pass all exams in graduating courses in order to receive a diploma, and entry into tertiary education is based on grades received in upper secondary school. If a student fails an exam in their final year of upper secondary school, they must retake the exam before they can receive a diploma and continue onto higher education, requiring in most cases a one-year delay. Students who are enrolled in the academic track and fail an exam during their third year may apply to change to a vocational education track; however, this would require in most cases at least 2 years of further education and apprenticeship, and as not every school offers all tracks, students may need to move large distances to attend the desired vocational track.

If students choose to not retake the exam, and therefore to not receive an upper secondary diploma, they may enter the labor market without any certification, although this would often result in working in low-wage jobs. Otherwise, these students may choose to enroll in the military service, which would provide an alternative to attending tertiary education and a decent living wage, while also providing an opportunity to retake exams from upper secondary school. Students who do not retake exams may furthermore be admitted to tertiary education in a quota where application score requirements are lower. Students who fail an exam may therefore have to spend more time and resources to enter their study of choice, through retaking more exams or gaining additional points (e.g., for military service or age). An overview of the Norwegian education system is provided in Figure 1.

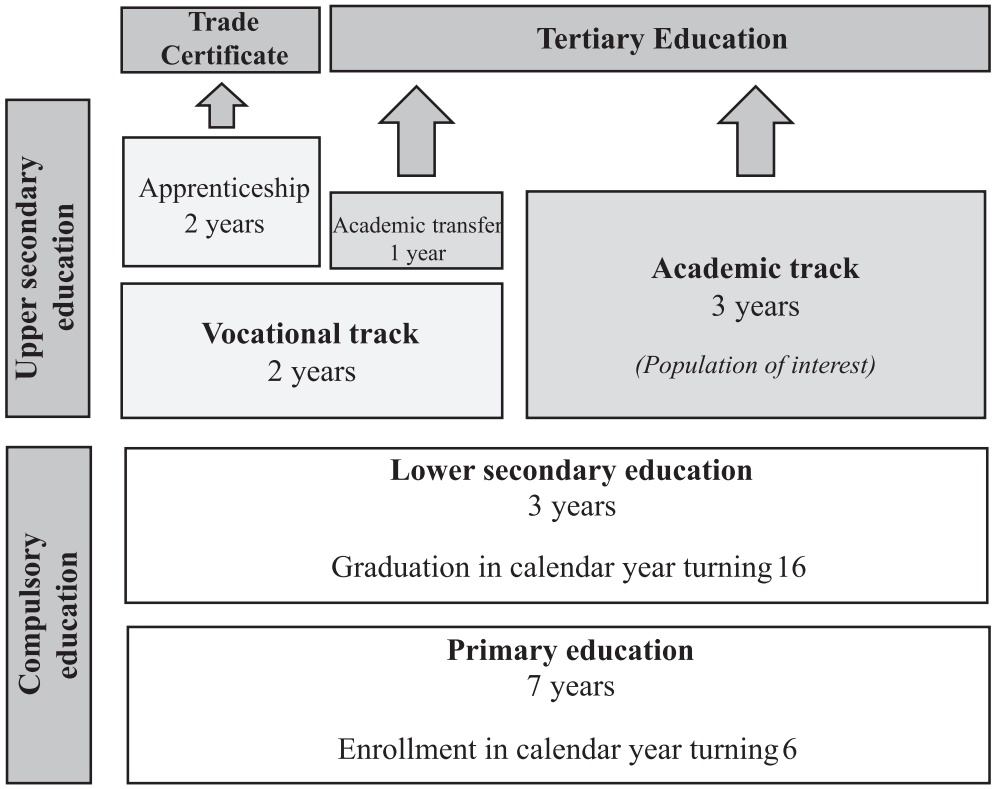

FIGURE 1 Norwegian education system. Adapted from Falch et al. (2014).

Primary healthcare system

In Norway, 85% of total healthcare expenditure comes from public financing, with all residents entitled to medical and healthcare services and all individuals regardless of citizenship or resident status entitled to acute emergency services (Saunes et al., 2020). All residents in Norway are entitled to a primary care physician, resulting in a high level of coverage with 98% of the population on a primary care physician's list (Norwegian Directorate of Health, 2020). Primary care physicians play a key role in the provision of healthcare services and are responsible for providing primary diagnoses, prescribing medications, and referring patients to specialist healthcare when needed. Primary care physicians act as gatekeepers to specialist services and only primary care physicians and ambulance services can refer patients to hospital admission or emergency hospital care (Saunes et al., 2020). This system is designed to ensure continuity of care, therefore an individual's first point of contact within the healthcare system is usually through the primary care physician, or for more urgent cases, through the on-call physician at emergency centers (Saunes et al., 2020).

The psychological diagnoses used in this study come from information on primary care visits, and the specific diagnoses included in our analysis are those related to stress and anxiety, depression, sleep disturbances, and substance abuse. These are most likely to be related to failing a high-stakes exam through academic stress and disappointment, as well as the use of unhealthy coping mechanisms. We chose to be inclusive in the diagnostic codes included, to allow for variations in coding behaviors of primary care physicians and a wide range of responses to failing an exam, while removing those which are theoretically unrelated to the exam. A detailed description of the diagnostic data is provided in the Measures section and further in Supporting Information Appendix A.

MATERIALS AND METHODS

Data sources and study population

The study population included all students on an academic track in the third year of upper secondary school who scored either a 1 (failed) or 2 (passed) on the Norwegian first language written exam from 2006 to 2018. The population is limited to students with the two lowest grades on the exam as both grades signal relatively weak exam performance. We selected 2006 as the first exam year in our sample as this is the first year that data on primary care diagnoses are available. Individual-level data were obtained through the linkage of three national registries containing information on the entire population residing in Norway using unique personal identifiers. Registries used include the Norwegian Population Register, the National Education Database, and the National Health Insurance Scheme

(KUHR). This registry linkage provided information on background characteristics, upper secondary grades and exam scores, lower secondary grades and GPA, parental education levels, primary care physician visits and diagnoses, graduation from upper secondary education, and enrollment in tertiary education programs.

To define the sample, we started with all students who took a Norwegian first language written exam and who were enrolled in an academic track and in their third year of upper secondary from school years 2005 to 2006 until 2017 to 2018. Students missing lower secondary school GPA were excluded from the study population. These are students who have received less than half of the grades (usually due to absence or recent migration to Norway) and are exempted from GPA-based competition for upper secondary schools and programs. Additionally, students missing a teacher-assessed grade for the written first language Norwegian course in upper secondary school were excluded from the sample. The sample was also restricted to exclude those 22 years and older at exam, as few students are enrolled in upper secondary school past this age. The government provides the right for all students to attend upper secondary school within 5 years of graduating compulsory school. As nearly all individuals finish compulsory school in the year they turn 16, students 22 years and older at the time of taking the exam would most likely not be covered under this right. These students are therefore likely attending adult education schools and are not comparable to the majority of upper secondary students. Individuals aged 22 years and older comprised slightly more than 1% of the sample before exclusion. This left 349,633 students receiving all grades (1–6) on the exam; of those, our sample was restricted to the 65,901 students who received a 1 or 2 grade. The sample selection is shown in Table 1.

Measures

The main exposure variable was failing the Norwegian first language written exam (coded with a dummy variable). With few exceptions, all students are required to take this exam in their final year of upper secondary school and must pass this exam to graduate and be eligible for entry into tertiary education programs. Only students enrolled in the academic tracks were selected for the analysis, as these tracks specifically prepare students for entry into tertiary education, as opposed to vocational tracks which focus on preparing students for entry into specific occupations following graduation.

The main outcome variable of interest was receiving a psychological diagnosis from a primary care physician after the exam and within the following year (dichotomized). This was obtained through reimbursement data in the National Health Insurance Scheme (KUHR) database. Primary care physicians are required to submit at least one International Classification of Primary Care version 2 (ICPC-2) code per patient visit to receive reimbursement, and patients are often required to receive a referral from primary care physicians before receiving specialist healthcare services. While the outcome measure of interest was receiving a psychological diagnosis, nine ICPC-2 psychological codes were excluded from the analysis as they are not considered to be theoretically related to the effects of failing an exam, leaving 34 ICPC-2 diagnoses in the analysis (World Organization of National Colleges and Academies, 2005). The ICPC-2 diagnostic codes are provided in Supporting Information Appendix A. Graduation from upper secondary school and enrollment in tertiary education were also explored for the following timeframes: within the calendar year of the exam date, within the next calendar year following the exam, and within 5 years of the exam.

TABLE 1 Sample selection.

Sample selection	*N* (%)
Enrolled in third year academic track between 2005/2006 and 2017/2018	409,528 (100)
Received a Norwegian written exam grade	364,391 (89.0)
Not missing the teacher-assessed written Norwegian grade	357,076 (87.2)
Not missing lower secondary grade point average	353,708 (86.4)
Age at exam less than 22	349,633 (85.4)
Received a 1 or 2 grade on exam	65,901 (16.1)
Unmatched sample	65,901 (100)
Received a 2 (passed)	56,812 (86.2)
Received a 1 (failed)	9089 (13.8)

Empirical strategy

Students who fail exams may differ from those who barely pass in several respects beyond their exam performance, for example, in socioeconomic background or prior academic performance. To allow for a meaningful comparison between these groups, we therefore utilized propensity score matching (PSM) methods (Rosenbaum & Rubin, 1983). This method creates two groups—a "treated" group (those who failed) and a "control group" (those who barely passed)—which exhibit similar distributions on observed covariates, differing only on the treatment status: passing or failing the exam. As matching on all observed covariates by using exact matching methods may result in the loss of observations and poor matches, PSM matches on the propensity score, or the predicted probability of failing the exam (vs. passing), conditional on observed characteristics.

Propensity scores were estimated using logistic regression conditional on the following characteristics: legal registered gender (dummy), age (continuous, measured in years), lower secondary GPA (third-degree polynomial), teacher-assessed grade (discrete), exam year

(discrete), immigration background (indicator), mother's highest education level (indicator), and father's highest level of education (indicator). Legally registered gender is the gender which the individual is registered by the government. Individuals may apply to legally change their gender, and such change would result in an update of administrative data to reflect this. Lower secondary school GPA was standardized for each cohort and parental education levels were measured at age 16 and classified into the following categories: lower secondary and below, upper secondary, bachelor's level, master's level and above, and missing. The teacher-assessed grade is a final grade which appears on the student's diploma and is a measure of the student's performance in the exam subject at the end of term. In contrast to the exam grade, this grade is set by the teacher themselves, rather than being anonymously set by teachers from another region. Similar to the exam grade, the teacher-assessed grade is meant to be an assessment of the student's competence and teachers are monitored on how well these two grades match. Immigration background was categorized as a combination of immigration status (child of Norwegian-born parent(s), child of immigrant parent(s), or immigrant) and the country background of either the parents or the individual themselves (Norwegian, Western, or Non-Western). Information regarding the ethnic and racial characteristics of the sample is not available as these data are not collected in administrative registers by Statistics Norway. Demographic characteristics of the sample are presented in Table 2.

Subjects were matched using 1-to-1 nearest neighbor without replacement. The matching strategy ensures that treated participants are matched with the control participant with the nearest propensity score, within a specified distance (caliper) which improves matching quality and enforces common support (Leite, 2017). To avoid compression around 0 and 1, we used logit propensity scores (Austin, 2011). Covariate balance before and after matching was assessed using standardized mean differences (SMDs) for each covariate, and variance ratios for continuous variables. SMDs allow for comparison across continuous and dichotomous variables. As ideally all measures of central moments of the covariates are the same between the treated and control groups, variance ratios for continuous variables were also used to assess balance after matching. The covariates are considered balanced if the SMD was less than 0.10, and the variance ratio was less than 2 (Austin, 2009).

To estimate the association between failing a graduating exam and mental health, odds ratios (OR) for psychological diagnosis (dichotomized) were estimated using logistic regression with a dummy variable for failing (ref. passing) the exam in the propensity score matched sample. To examine the rate of psychological diagnoses, rather than solely the presence or absence of a diagnosis, incidence rate ratios (IRRs) were calculated using negative binomial regression for the number of diagnoses (count). Examining both the extensive margin of psychological diagnoses through ORs and the intensive margin through IRRs allowed us to examine whether a relative increase in psychological diagnoses was due to not only the same individuals receiving a higher rate of diagnoses, but whether there was an increase in the number of individuals receiving any diagnosis after the exam. Both measures of psychological diagnoses were assessed during the semester following the exam and until the end of the following calendar year (up to 1.5 years). These two measurements provided similar estimates, and the IRRs are available in Supporting Information Appendix B. We also calculated marginal effects estimates which are presented in Supporting Information Appendix C.

We used subgroup analyses to estimate differences by gender, teacher-assessed grade received, immigration background, and parental income at age 16. ORs were also estimated for graduating from upper secondary education and enrollment in tertiary education within the same calendar year of the exam, within the next calendar year, and within 5 years of the exam.

To examine possible differences in psychological diagnosis prevalence between the two groups not due to the exam, we used a pre- and post-trend analysis. For the trends analysis, a sub-sample was created by restricting the matched sample to students living in Norway for the 3 years prior to the exam. This is to ensure that data on psychological diagnoses were available for the entire analytical sample. The sample was additionally restricted to students during exam years 2009–2017 due to data restrictions. For each 1-year period, starting 3 years prior to the exam until 2 years following the exam, ORs were estimated for psychological diagnoses (dichotomized) for the treated group (failed), using the matched control group (passed) as the reference.

All analyses were adjusted for the following covariates: gender, age, standardized GPA (third-degree polynomial), teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Logistic and negative binomial regression equations are provided in Supporting Information Appendix D. As suggested by Nguyen et al. (2017), "double-adjustment" by including covariates used for matching in the regression analyses helps to reduce bias in estimates due to covariate imbalance and is robust to misspecifications in the propensity score model, as compared to using the propensity score alone.

We conducted a sensitivity analysis to assess for residual bias using unrelated diagnoses as a placebo outcome test. Additionally, placebo treatment tests were also completed using two "fake" treatment groups with students who received a 2 ("treated") and a 3 ("control") as well as students who received a 3 ("treated") and a 4 ("control") on the exam. These groups were matched using the same matching procedure, replicating the main analysis in both matched samples. Furthermore, to assess for sensitivity of the propensity score model to

TABLE 2 Descriptive statistics.

	All students	Exam grades 3–6	Exam grade 2	Exam grade 1	Matched exam grade 2	Matched exam grade 1	Standardized mean difference
Girls (*M* (SD))	0.56 (0.50)	0.58 (0.49)	0.44 (0.50)	0.36 (0.48)	0.36 (0.48)	0.36 (0.48)	0.0074
Age (*M* (SD))	19.05 (0.26)	19.04 (0.24)	19.10 (0.32)	19.13 (0.40)	19.12 (0.39)	19.13 (0.40)	0.0303
Birth month (*M* (SD))	6.29 (3.38)	6.27 (3.38)	6.37 (3.38)	6.42 (3.37)	6.43 (3.36)	6.41 (3.37)	−0.0064
Standardized GPA *z*-score (*M* (SD))	0.63 (0.69)	0.77 (0.63)	0.09 (0.63)	−0.25 (0.67)	−0.23 (0.65)	−0.24 (0.66)	−0.0133
Standardized GPA of peers *z*-score (*M* (SD))	0.63 (0.21)	0.63 (0.21)	0.62 (0.22)	0.61 (0.23)	0.61 (0.21)	0.61 (0.23)	0.0121
Missing (*n*)	1176	916	213	47	34	47	0.0014
Student cohort size (*M* (SD))	321 (176)	325 (174)	306 (181)	298 (200)	305 (191)	298 (199)	−0.0357
Missing (*n*)	1580	1072	395	113	67	112	0.0050
Centrality index of municipality (*M* (SD))	2.71 (1.27)	2.68 (1.26)	2.83 (1.27)	2.86 (1.31)	2.79 (1.28)	2.87 (1.31)	0.0550
Missing (*n*)	1117	823	240	54	40	54	0.0016
Birth order (*M* (SD))	1.85 (0.96)	1.84 (0.95)	1.89 (1.01)	1.92 (1.07)	1.91 (1.04)	1.91 (1.06)	0.0064
Number of children in household (*M* (SD))	1.95 (0.95)	1.94 (0.94)	1.96 (1.00)	2.00 (1.05)	1.96 (1.01)	2.00 (1.05)	0.0388
Missing (*n*)	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	<0.5%	0.0000
Psychological diagnoses (*M* (SD))	0.17 (0.95)	0.16 (0.94)	0.18 (1.00)	0.24 (1.15)	0.19 (0.95)	0.24 (1.15)	NA
Psychological diagnosis (share)	6.82%	6.61%	7.47%	9.37%	7.80%	9.44%	NA
All diagnoses (*M* (SD))	2.07 (2.95)	2.05 (2.91)	2.14 (3.10)	2.26 (3.38)	2.13 (3.14)	2.27 (3.38)	NA
	Median (IQR)	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	**Median (IQR)**	
Annual household income at age 16 (NOK)	710,631 (387,420)	725,553 (394,498)	658,266 (352,721)	620,585 (335,379)	621,924 (337,484)	621,613 (334,666)	0.0059
Missing (*n*)	19,830	16,367	2992	471	464	468	0.0004
	%	%	%	%	%	%	
Region of residence (share)							
Northern Norway	8.44	7.96	10.25	12.23	9.42	12.29	0.0287
Eastern Norway	50.68	50.65	50.85	50.58	53.32	50.48	−0.0285
Western Norway	26.39	27.01	23.98	22.06	23.69	22.08	−0.0161
Southern Norway	5.79	5.74	5.96	6.07	5.56	6.08	0.0052
Trøndelag region	8.69	8.63	8.96	9.05	8.01	9.07	0.0106
Teacher-assessed grade (share)							
1	<0.5	<0.5	1.25	6.93	6.03	6.32	0.0029
2	7.60	3.37	23.06	43.10	44.05	43.36	−0.0069
3	27.02	22.55	47.75	36.85	37.14	37.10	−0.0003
4	37.72	41.38	23.75	11.05	11.10	11.12	0.0002

(Continues)

TABLE 2 (Continued)

	%	%	%	%	%	%	
5	23.27	27.82	4.00	1.90	1.50	1.92	0.0042
6	3.92	4.78	<0.5	<0.5	<0.5	<0.5	−0.0001
Exam year (share)							
2006	5.35	5.50	4.77	4.31	4.37	4.33	−0.0003
2007	6.08	6.01	6.33	6.65	6.71	6.64	−0.0008
2008	6.71	6.59	7.19	7.42	7.89	7.46	−0.0043
2009	7.44	7.15	8.42	10.25	9.82	10.23	0.0041
2010	7.74	7.32	9.44	10.01	9.78	10.02	0.0023
2011	7.88	7.36	9.42	14.79	15.43	14.64	−0.0080
2012	7.70	7.45	8.65	9.57	9.78	9.54	−0.0024
2013	8.05	8.21	7.51	6.58	7.01	6.61	−0.0040
2014	8.26	8.22	8.40	8.31	8.10	8.32	0.0022
2015	8.60	8.64	8.45	8.03	7.88	8.07	0.0019
2016	8.70	8.90	8.14	6.21	6.02	6.24	0.0022
2017	8.74	9.12	7.52	4.48	4.08	4.50	0.0042
2018	8.76	9.53	5.75	3.40	3.14	3.42	0.0029
Immigration background (share)							
Norwegian-born parents: Norway born	75.08	76.25	70.76	65.60	69.26	65.83	−0.0342
Norwegian-born parents: western born	7.00	7.23	6.09	5.41	5.09	5.44	0.0035
Norwegian-born parents: non-western born	1.68	1.66	1.78	1.87	1.62	1.88	0.0027
Immigrant: western origin	1.05	0.96	1.46	1.39	1.23	1.40	0.0017
Immigrant: non-western origin	2.71	1.97	5.41	8.81	7.64	8.59	0.0094
Immigrant parents: western origin	<0.5	<0.5	<0.5	<0.5	<0.5	<0.5	0.0004
Immigrant parents: non-western origin	4.01	3.40	6.40	8.15	7.12	8.11	0.0099
1 Norwegian-born parent: Western origin	5.78	5.93	5.14	5.12	4.51	5.13	0.0062
1 Norwegian-born parent: non-western origin	2.36	2.27	2.65	3.26	3.19	3.24	0.0004

TABLE 2 (Continued)

	%	%	%	%	%	%	
Mother's education (share)							
Lower secondary and lower	13.23	11.55	19.95	23.73	22.73	23.63	0.0090
Upper secondary	35.63	34.52	40.41	40.37	42.24	40.44	−0.0181
Bachelor level	40.85	42.81	33.01	28.73	28.91	28.84	−0.0007
Master level and higher	9.41	10.50	4.85	3.97	3.62	3.97	0.0034
Missing	0.88	0.62	1.79	3.20	2.49	3.12	0.0063
Father's education (share)							
Lower secondary and lower	12.48	11.19	17.65	20.62	19.84	20.55	0.0071
Upper secondary	42.37	41.23	47.31	46.87	49.22	46.99	−0.0224
Bachelor level	27.12	28.38	22.13	19.24	18.98	19.29	0.0031
Master level and higher	15.73	17.31	9.14	7.48	6.89	7.50	0.0061
Missing	2.30	1.90	3.78	5.79	5.06	5.67	0.0061
Graduation from upper secondary (share)							
Within exam year	85.89	90.97	71.43	18.01	59.23	18.10	NA
Within 1 year (2006–2017)	90.18	94.23	78.67	42.13	67.24	42.34	NA
Within 5 years (2006–2013)	93.84	96.73	86.52	62.37	77.31	62.72	NA
Enrollment in tertiary education (share)							
Within exam year	35.08	38.45	23.53	2.01	16.83	2.02	NA
Within 1 year (2006–2017)	68.15	73.21	50.77	26.05	39.75	26.19	NA
Within 5 years (2006–2013)	90.06	93.99	78.50	56.36	67.40	56.66	NA
N 2006–2013	199,095	157,696	35,075	6324	6390	6269	NA
N 2006–2017	319,009	256,686	53,543	8780	8743	8717	NA
N	349,633	283,732	56,812	9089	9026	9026	NA

Note: Gender, age, birth month, standardized grade point average (GPA), parents' income and immigrational background are used in the matching, while psychological diagnosis, graduation from upper secondary school, and enrollment in tertiary education are outcome variables. Other available variables are included in the table to expose potential unbalance in the matched sample. Region of registered residence, teacher-assessed grade, exam year, immigration background, mother's education level, father's education level, graduation from upper secondary education and enrollment in tertiary education are shares of column totals. Income measured in Norwegian Krone (NOK).

the selection of covariates, we conducted a sensitivity analysis by repeating the matching procedure and main analysis while removing one of the matching variables for all eight covariates, and the results of this are presented in Supporting Information Appendix F, Table F1. We also conducted sensitivity analyses for the matching model specification by including multiple control units and the use of replacement in the matching, as well as by using propensity score weighing methods. The results of these sensitivity analyses are presented in Supporting Information Appendix F, Tables F2 and F3. Finally, a sensitivity analysis was completed by including additional covariates to the matching strategy. The covariates added were region of residence (categorical), centrality index (continuous), GPA average of peers (continuous), class cohort size (categorical), number of children in household (categorical), and lowest-income quintile (indicator). The results remained robust to the inclusion of additional matching covariates and the sensitivity analysis is presented in Supporting Information Appendix F, Table F4. The analyses regarding the association between high-stakes exam failure and psychological diagnoses should be considered relatively exploratory due to the limited prior research on the topic. The analyses regarding educational outcomes, however, can be considered relatively confirmatory due to prior research on high-stakes exam failure and further education and dropout in the United States and the United Kingdom. All analyses were conducted using R version 4.1.2 and matching was done with the MatchIt package in R (R Core Team, 2021; Stuart et al., 2011).

RESULTS

Descriptive statistics

Starting with the full sample of students who passed and failed the exam (n=65,901: passed (n)=56,812, failed (n)=9089), we applied 1-to-1 nearest neighbor PSM within caliper. This resulted in the analytical sample consisting of 18,052 students, with 9026 students who received a 1 (failed), and 9026 students who received a 2 (passed). Table 2 shows the descriptive statistics of students by exam grade both before and after PSM, along with SMDs for covariate distance between the matched groups. Before matching, 55.9% of students receiving a 2 were boys, compared to 63.7% of those who failed the exam. The shares receiving the lowest exam grade varies greatly by year from 3.4% to 14.8%, probably reflecting variation in test difficulty. Shares receiving exam grade 2 co-vary with this pattern yet are somewhat more stable. The covariate with the largest difference between the two groups before matching was the standardized lower secondary school GPA. Those who passed the exam had an average lower secondary school GPA of slightly above the mean (z-score=0.091), while those who failed the exam had on average a GPA of 0.249 standard deviations below the mean.

After matching, the gender distribution between the two groups was improved with boys accounting for 64.4% and 63.6% of those who received an exam grade of 2 and 1 in the matched sample, respectively. The absolute difference in the proportion of students who immigrated from a non-Western country between those who failed and those who passed also improved after matching, changing from a 3.4% difference to a 0.9% difference. Covariate balance across the groups was also greatly improved for the teacher-assessed course grades, and standardized lower secondary GPA. All other covariates were balanced according to SMD and variance ratios after matching, shown in Figure 2. The covariate with the largest remaining SMD after matching was those with two Norwegian-born parents who were born in Norway (0.073). This was within the 0.10 SMD threshold for imbalance, although it may reflect slight heterogeneity remaining in the matched sample.

Matching also improved balance between the groups for household income at age 16, a variable which was not included in the matching covariates but reflects the socioeconomic background of the groups. Additionally, matching improved the balance for birth month, standardized GPA average of peers, student cohort size, centrality index of municipality, birth order, number of children in the household, and registered region of residence between those who failed and passed the exam. These variables, while not included in the matching procedure, were all considered balanced with an SMD of less than ±0.10 after matching. The distributions of propensity scores before and after matching are provided in Supporting Information Appendix E.

Exam failure and psychological diagnoses

Table 3 shows the OR for psychological diagnoses in the matched sample. The OR describes the relative change in odds of receiving at least one diagnosis in students who failed compared to those who passed within the following calendar year post-exam. Students who failed the exam were at 1.21 (95% CI: 1.09–1.35, Z=3.58, p<.01) times higher odds of receiving at least one psychological diagnosis compared to those who passed with a 2.

Table 4 shows the OR for psychological diagnoses within one calendar year of the exam split by gender, teacher-assessed grade, immigration background, and parental income at age 16. We found no significant effects of failing the exam on girls (OR 1.11, 95% CI: 0.95–1.30, Z=1.30, p=.192). In the male subgroup, boys who failed were at 1.31 (95% CI: 1.13–1.51, Z=3.62, p<.01) times the odds of receiving a diagnosis compared to boys in the control group. The effect of failing was larger for those who had received a high teacher-assessed course grade

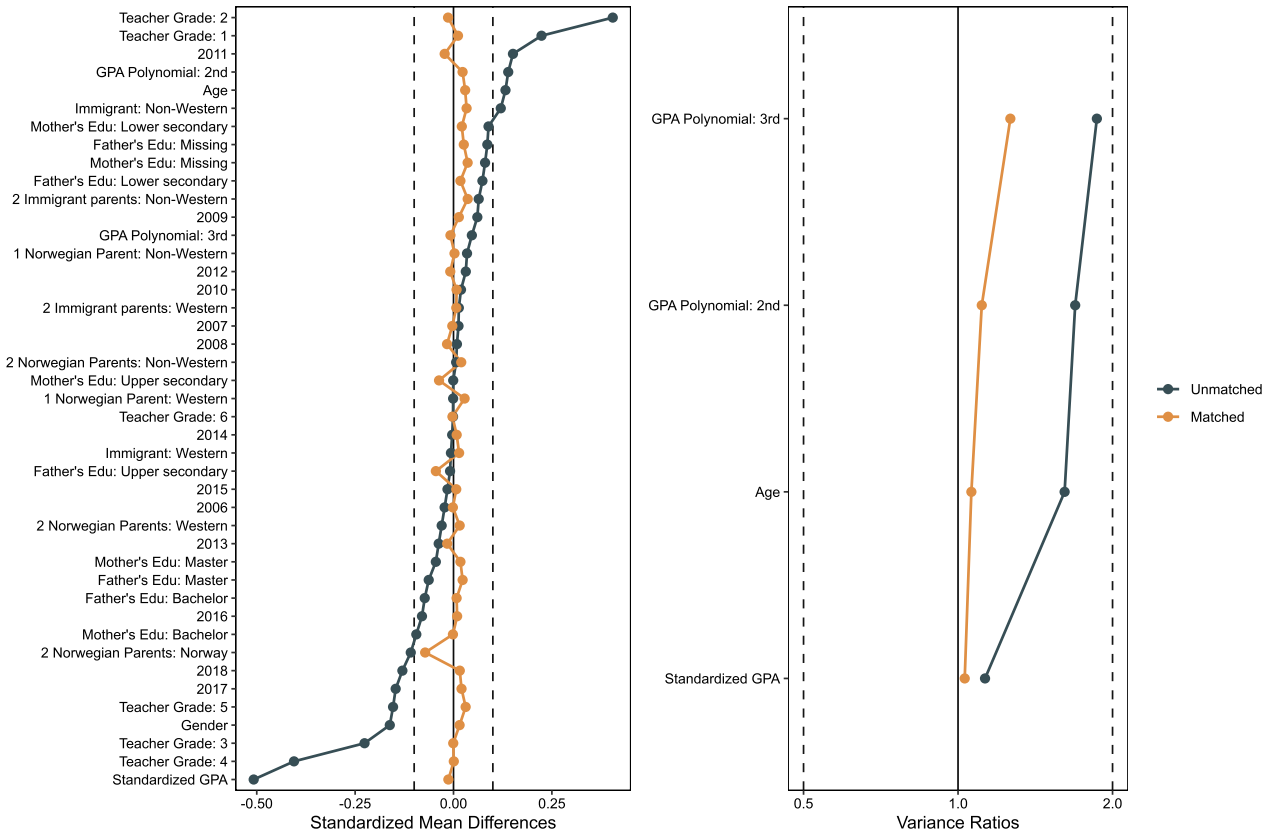

FIGURE 2 Covariate balance. Standardized mean differences and variance ratios of matching covariates pre- and post-matching. GPA, grade point average.

TABLE 3 Effect of exam failure on psychological diagnoses within one calendar year.

Matched sample	
	OR (95% CI)
Failed	1.21 (1.09–1.35)***
Passed (ref)	1.00
N	18,052

Note: Adjusted for gender, age, standardized grade point average third-degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Confidence intervals using White-robust standard errors.

***$p < .01$.

(4–6). Among these students, those who failed were at 81% increased odds of receiving a psychological diagnosis (95% CI: 31–150, $Z = 3.58$, $p < .01$) compared to those who passed the exam with high teacher grades. Students who failed the exam and had received lower teacher-assessed grades (1–3) had 15% increased odds of receiving a diagnosis (95% CI: 3–29, $Z = 2.42$, $p < .05$) compared to those who passed with a similar teacher grade.

Among those with at least one foreign-born parent, those who failed the exam were at 25% increased odds of receiving a psychological diagnosis (95% CI: 2–54, $Z = 2.16$, $p < .05$) compared to those who passed. This was similar to those with two Norwegian-born parents, as those who failed were at 21% increased odds (95% CI: 7–37, $Z = 3.00$, $p < .01$) compared to those who passed. For students whose household income was in the lower 50% of all students at age 16, those who failed were at a 19% increased odds of receiving a psychological diagnosis (95% CI: 4–36, $Z = 2.53$, $p < .05$) compared to those who passed, while those who failed with household income at age 16 in the upper 50% of students were at a 24% increased odds compared to those who passed (95% CI: 3–49, $Z = 2.27$, $p < .05$).

Pre- and post-exam psychological diagnoses

We conducted a pre- and post-trends analysis to better understand whether the increase in psychological diagnoses following the exam was instead due to underlying differences, such as differential time trends, between the treatment and control groups. If the two groups differ in such a way, our previous estimates using a singular time frame would fail to account for these differential trends, and therefore may be biased by capturing more than the impact of failing the exam. To address this, we calculated ORs and the mean number of psychological diagnoses in each 365-day period for the treatment and control groups starting 3 years prior to the exam until 2 years after. If the

TABLE 4 Stratified analysis of the effect of exam failure on psychological diagnoses within one calendar year.

	Gender		Teacher-assessed grade	
	Girls	Boys	Low grades (1–3)	High grades (4–6)
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Failed	1.11 (0.95–1.30)	1.31 (1.13–1.51)***	1.15 (1.03–1.29)**	1.81 (1.31–2.50)***
Passed (ref)	1.00	1.00	1.00	1.00
N	6499	11,553	15,705	2347
	Immigration background		**Household Income at age 16**	
	1 Foreign-born parent	2 Norwegian-born parents	Below median	Above median
	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
Failed	1.25 (1.02–1.54)**	1.21 (1.07–1.37)***	1.19 (1.04–1.36)**	1.24 (1.03–1.49)**
Passed (ref)	1.00	1.00	1.00	1.00
N	4593	13,459	10,330	6790

Note: Adjusted for gender, age, standardized grade point average third degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Stratifying variable not adjusted for in analysis. Excluding 932 individuals with missing parental income at age 16 in income stratified analysis. Confidence intervals using White-robust standard errors.

**$p < .05$; ***$p < .01$.

trends between our two groups move in parallel prior to the exam, then under plausible assumptions, the control group would provide a reasonable counterfactual, and we would expect that the treatment group would continue to follow this trend if not for failing the exam.

The time periods used capture one school year each, from July until June, with the "exam year" covering the third year of upper secondary school ending with the written Norwegian exam in June. The matched sample was restricted to individuals who lived in Norway for the 3 years prior to the exam and to students taking the exam between 2009 and 2017 due to data restrictions ($n = 14{,}063$). Figure 3 shows the average number of psychological diagnoses for each time period in both the students who passed and failed the exam. The average number of diagnoses is slightly higher among the students who failed the exam; however, both groups follow parallel trends throughout the pre-exam period. Following the exam, students who failed experienced a slight increase in the average number of diagnoses, while the students who passed the exam experienced a slight drop in the average number of diagnoses.

Figure 4 shows the ORs for receiving at least one psychological diagnosis for the restricted sample of students who failed the exam in each 1-year period compared to the matched control group. The ORs for each pre-exam period were not statistically significant. However, in the first year following the exam, students who failed had 1.33 (95% CI: 1.10–1.61, $Z = 2.96$, $p < .01$) times the odds of receiving a psychological diagnosis compared to students with the lowest passing grade. Following this increase, the OR for 2 years post-exam returns to statistically insignificant levels seen in the pre-exam periods.

Exam failure and educational outcomes

Table 5 shows the ORs for graduation from upper secondary education and enrollment in tertiary education within the same calendar year as the exam, within the next calendar year following the exam, and within 5 years of the exam. Perhaps unsurprisingly, students who failed the exam were significantly less likely to graduate from upper secondary education (OR = 0.11, 95% CI: 0.10–0.12, $Z = -55.95$, $p < .01$) or enroll in tertiary education (OR = 0.09, 95% CI: 0.08–0.11, $Z = -29.42$, $p < .01$) within the same year of the exam compared to the matched control group. While the difference between the groups was somewhat reduced when considering outcomes within 1 year of the exams, students who failed the exam still had 71% (95% CI: 68–73, $Z = -35.15$, $p < .01$) lower odds of graduating from upper secondary education compared to those who passed. This persisted for the 5-year period following the exam, by which time students who failed experienced a 57% (95% CI: 53–61, $Z = -19.69$, $p < .01$) reduction in the odds of graduating compared to students who passed with the lowest passing grade.

For enrollment in tertiary education, students who failed had a 51% (95% CI: 48–55, $Z = -20.51$, $p < .01$) reduction in the odds of enrolling within 1 year of the exam. This effect was slightly attenuated during the 5 years post-exam period, although students who failed still had a 44% (95% CI: 39–48, $Z = -14.16$, $p < .01$) reduction in the odds of enrolling compared to passing students. To further understand the long-term outcomes of the students who fail a high-stakes exam, we examined the proportion who were considered not in employment, education, or training (NEET) for the first 5 years after the exam. In Figure 5, we can see that among the matched population, about 12.5% of those who failed were NEET 1 year after

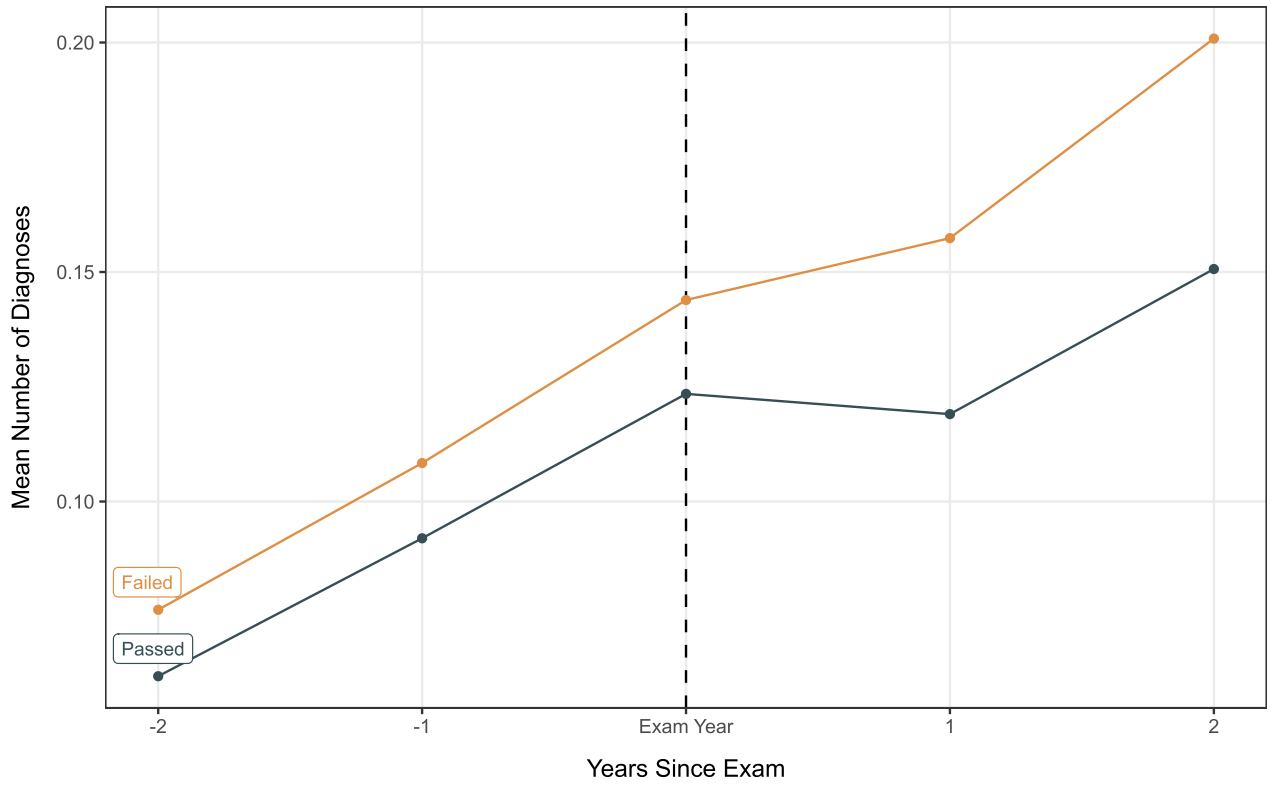

FIGURE 3 Mean number of pre- and post-exam psychological diagnoses. Matched sample restricted to those living in Norway for 3 years prior to the exam and exam years 2009–2017 ($n = 14{,}063$). Each year captures one 365-day period from July to June.

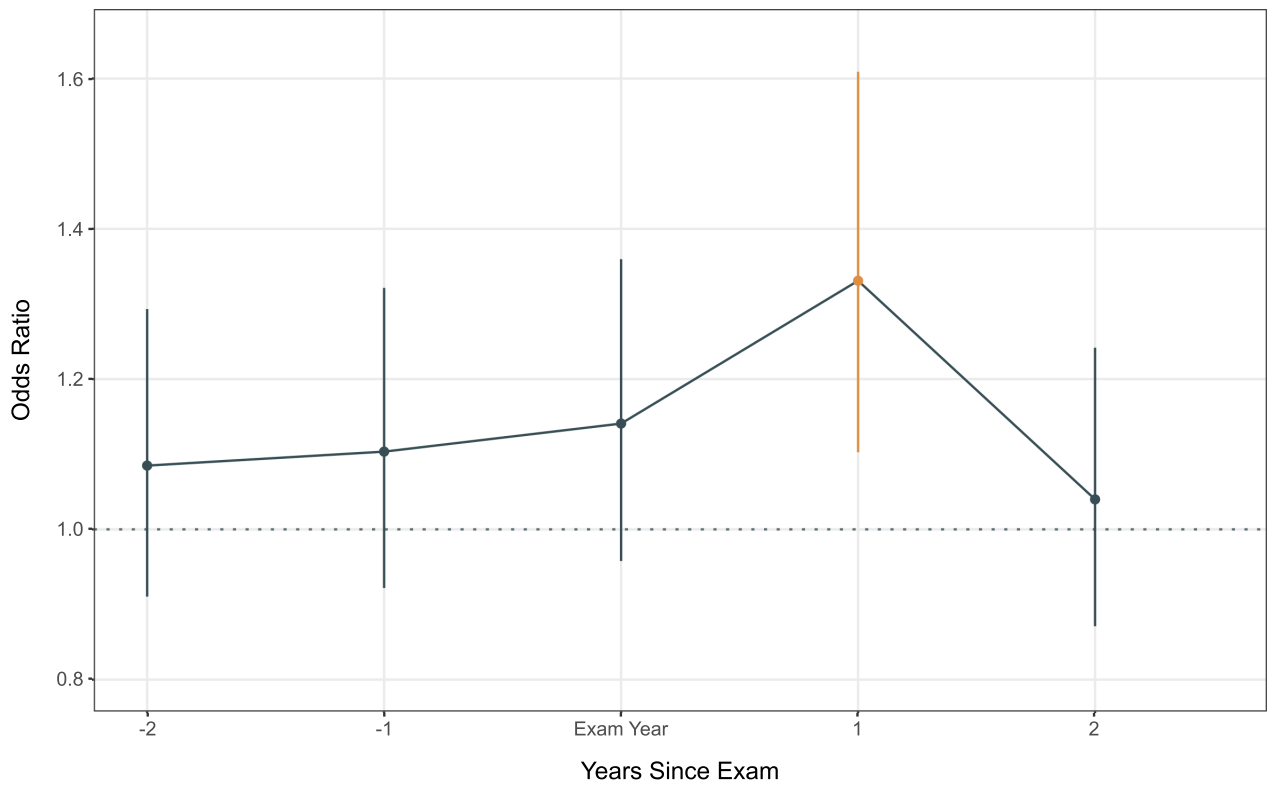

FIGURE 4 Odds ratios for pre- and post-exam psychological diagnoses. Matched sample restricted to those living in Norway for 3 years prior to the exam and exam years 2009–2017 ($n = 14{,}063$). Each year captures one 365-day period from July to June. All ORs adjusted for gender, age, standardized grade point average third-degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Bars show 95% confidence intervals on point estimates using White-robust standard errors in yellow (significant) and black (insignificant).

TABLE 5 Effect of exam failure on educational outcomes.

	Graduating within exam year	Graduating within 1 year	Graduating within 5 years
	OR (95% CI)	OR (95% CI)	OR (95% CI)
Failed	0.11 (0.10–0.12)***	0.29 (0.27–0.32)***	0.43 (0.39–0.47)***
Passed (ref)	1.00	1.00	1.00
	Enrolled within exam year	**Enrolled within 1 year**	**Enrolled within 5 years**
Failed	0.09 (0.08–0.11)***	0.49 (0.45–0.52)***	0.56 (0.52–0.61)***
Passed (ref)	1.00	1.00	1.00
N	18,052	17,460	12,659

Note: Adjusted for gender, age, standardized grade point average third degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Confidence intervals using White-robust standard errors.
***$p<.01$.

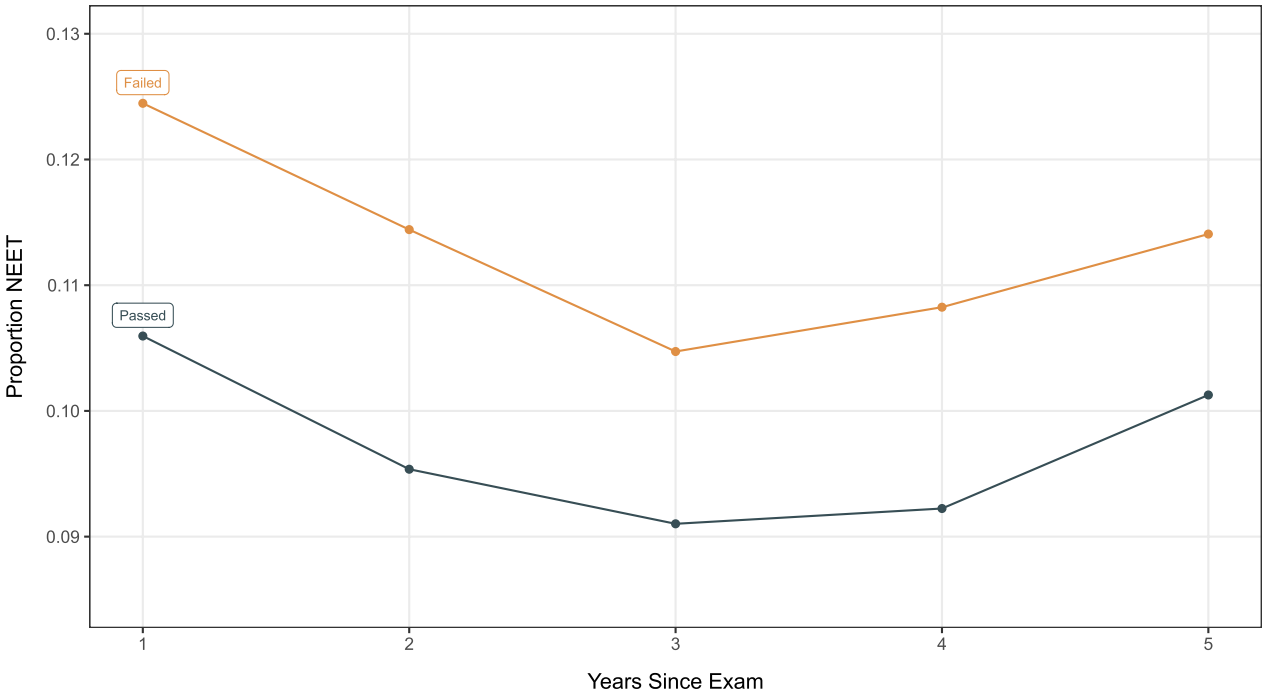

FIGURE 5 Proportion not in employment, education, or training (NEET) up to 5 years post-exam. Matched sample ($N=18,052$).

the exam, compared to about 10.5% of those who passed. After 5 years, about 11.5% of those who failed were considered NEET, compared to just above 10% of those who passed.

Sensitivity analyses

As potential residual selection bias may remain if the PSM strategy was not appropriate, we conducted a placebo outcome sensitivity analysis examining the effect of failing the exam on unrelated diagnoses (dichotomized and count). We would not expect failing the exam to have any impact on either the odds or rate of non-psychological diagnoses, respiratory diagnoses, neurological diagnoses, and skin diagnoses received during the semester following the exam and the following calendar year. These diagnoses capture both general healthcare seeking behavior, infectious, and noninfectious illnesses, which should not be impacted by failing the exam. However, if we are capturing a difference in the general propensity to seek medical care for the two groups rather than an effect of the exam on psychological distress, we would see significant differences in these unrelated diagnoses following the exam.

The results of this analysis are presented in Table 6. Students who failed the exam did not have a statistically significant difference in the odds of receiving any of the unrelated diagnoses, nor did they have a statistically significant difference in the rate of these diagnoses compared to those who passed. Although the results are consistent with the belief that exam failure is not related to these diagnoses, it is important to keep in mind that chronic stress has been linked to impaired immune function (Morey et al., 2015). Thus, individuals who fail an

TABLE 6 Sensitivity analysis for placebo outcomes.

	All non-psychological diagnoses		Respiratory diagnoses	
	Share with diagnosis (%)	**OR (95% CI)**	**Share with diagnosis (%)**	**OR (95% CI)**
Failed	60.65	1.03 (0.97–1.10)	22.81	0.99 (0.92–1.06)
Passed (ref)	59.77	1.00	22.98	1.00
	Neurological diagnoses		**Skin diagnoses**	
Failed	5.27	1.06 (0.93–1.21)	16.34	0.97 (0.90–1.05)
Passed (ref)	4.92	1.00	16.61	1.00
N	18,052			

Note: OR adjusted for gender, age, standardized grade point average third degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Confidence intervals using White-robust standard errors.

TABLE 7 Placebo treatment tests on mental health diagnoses.

Matched exam grade 2 versus 3		Matched exam grade 3 versus 4	
	OR (95% CI)		**OR (95% CI)**
Grade 2	1.04 (0.99–1.09)	Grade 3	0.98 (0.94–1.02)
Grade 3 (ref)	1.00	Grade 4 (ref)	1.00
N	100,644		160,318

Note: Adjusted for gender, age, standardized grade point average third degree polynomial, teacher-assessed grade, exam year, immigration background, mother's education level, and father's education level. Confidence intervals using White-robust standard errors.

exam may experience chronic stress and poor mental health, as well as worse physical health.

A placebo treatment test was also completed to assess the validity of failing the exam as the exposure. In this test, we used two "fake" treatment groups. We completed the same matching procedure for students who received a 2 ("treated") and a 3 ("control") on the Norwegian written exam and replicated the main analysis for receiving at least one psychological diagnosis and the rate of psychological diagnoses in the calendar year following the exam. We also did this same procedure using students who received a 3 ("treated") and a 4 ("control") on the exam. If a confounding factor such as a stressful life event affects both exam performance and mental health outcomes, without there being any causal link between the two, we would expect the event to also affect the likelihood of receiving a 2 rather than a 3, or a 3 rather than a 4. The results of this analysis are presented in Table 7. In both matched placebo samples, no significant effect was found for the odds of receiving at least one psychological diagnosis, nor the rate of psychological diagnoses following the exam. IRRs for both sensitivity analyses are provided in Supporting Information Appendix B. The results of further sensitivity analyses are provided in Supporting Information Appendix F.

DISCUSSION

Three key conclusions can be drawn from this study. First, failing on a high-stakes exit exam is associated with an increase in the odds of receiving at least one psychological diagnosis in the year following the exam compared to students who received the lowest passing grade. Second, students who failed the exam were significantly less likely to graduate from upper secondary education up to 5 years following the exam compared to the students who just passed, despite retake opportunities. Third, failing the exam resulted in a significantly reduced odds of enrolling in tertiary education in not only the first year following the exam, but up to 5 years post-exam.

In spite of the increasing levels of psychological distress among adolescent girls in Norway, our analysis found no significant effect of failing the exam on mental health diagnoses for girls and found a stronger and significant effect for boys, suggesting that boys may have been more impacted than girls by the exam failure. One reason for the difference may be that boys tend to mobilize more resources and perform better when the exam stakes are high (Azmat et al., 2016) and that negative emotions when failing such an exam therefore take a larger toll on boys' mental health. Another factor is that female adolescents are better at identifying trauma reactions than males and more likely to seek help (Haavik et al., 2017). Since girls are more likely to seek help from GP and specialized health services, this study may underestimate the gender difference in the impact of exam failure on mental health. It may also be the case that they avoid developing more severe impacts of exam failure on mental health since they seek more help from school nurses, youth health stations, and through their social network. Results should furthermore be interpreted within the Nordic context where women's rights and female participation in labor market has come further than in any other region. In Norway, girls have

higher career ambitions than boys and the country has the largest gender gap in girls' favor in expectations to complete tertiary education among 15-year-olds of all Organisation for Economic Co-operation and Development (OECD) countries (Borgonovi et al., 2018). This is also reflected in a large female advantage in school performance and educational attainment (Norwegian Ministry of Education and Research, 2019). The larger impact on mental health for males is particularly worrisome since 64% of students who fail the exam are male.

Point estimates also suggested a larger association for those students who received a high teacher-assessed course grade (4–6) in the exam course than students who received low teacher-assessed grades (1–3). Failing the exam may have acted as a larger shock to the students who received a high grade before the exam, as they may have expected to perform well on the exam. There do not appear to be large differences in the point estimates between those of different immigration background, nor between different household income levels at age 16.

Our results are partly in line with several previous studies examining the effects of high-stakes exam failure on dropout and enrollment in further education (Andresen & Løkken, 2020; Machin et al., 2020; Ou, 2010; Papay et al., 2010). Andresen and Løkken (2020) found that over 45% of marginal students who failed the exam remained without an upper secondary diploma at age 27. This is mirrored in our findings, as we found significant and long-lasting associations between exam failure and both graduation from upper secondary education and enrollment in tertiary education within 5 years of the exam. Many studies have found academic consequences of high-stakes exams in the United States and the United Kingdom, such as increased probability of dropout and lower likelihood of enrolling in tertiary education, particularly among low-income and minority students (Machin et al., 2020; Ou, 2010; Papay et al., 2010). Our point estimates in contrast show a slightly higher odds of receiving a psychological diagnosis among students with a household income above the median. One explanation for this may be the relatively low levels of income inequality in Norway, compared to the United States and the United Kingdom, although much of the previous research has focused on academic outcomes for low-income and minority students, and few studies have examined the psychological and stress-related outcomes resulting from high-stakes exam failure for these groups. Additionally, students from high-income backgrounds may have higher expectations of their own performance and may therefore experience a shock after failing, similar to students who received higher teacher-assessed grades.

Fewer studies have examined the mental health consequences of high-stakes exams. Kumandaş and Kutlu (2010) found an increase in fear of underachievement and test anxiety among students taking a high-stakes exam for placement into secondary education schools, and Högberg and Horn (2022) found a 12% increase in school-related stress as a result of high-stakes testing using pooled data from 31 European countries, including Norway. In South Korea, Wang (2016) found that students who ranked lower than expected on the college entrance exam experienced a higher likelihood of suicidal ideation and lower levels of life satisfaction following the exam. Similarly to our findings on psychological diagnoses, this effect appeared to be relatively short-term, dissipating 1 year post-exam.

Strengths and limitations

The strengths of this study lie in the rich register data used which cover the entire Norwegian population and include high-quality characteristics on family background and previous school performance. These register data allows us to provide estimates which we believe are externally valid for students failing an exam in the Norwegian upper secondary academic track. We were able to implement PSM to create two groups with similar distributions of the observed characteristics. We believe that the groups are comparable given the detailed characteristics included and that our sample consists of only students with similar background, academic performance, and who receive a 1 or 2 exam grade. As a result, we are able to estimate the association of high-stakes exam failure on mental health, dropout, and enrollment in tertiary education.

The results of this study should be interpreted within the context of some limitations. As with any observational study, we cannot eliminate the possibility of residual confounding, although we believe that due to the rich individual level data along with the PSM method, this is unlikely to completely account for the findings. Nevertheless, there may be unobserved characteristics that affect the treatment and control groups differentially and that become influential at the same time as the exam. As we used primary care diagnoses, we cannot account for those who may have minor or transient psychological symptoms, and those who have not sought healthcare. It may also be possible that, due to the potential stigma of seeking medical care for psychological illness, individuals may have visited their primary care physician for a different, but related, reason such as for headaches or muscle aches. This would likely not be captured in the primary care diagnoses unless mental health was discussed between the patient and the physician. Finally, as primary care physicians may refer patients to specialist healthcare following a psychological diagnosis, we are unable to follow these individuals after they no longer receive care through their primary care physician and transition into specialist healthcare services.

Future research and implications

As previous research has mainly focused on academic outcomes, many questions still remain on the

consequences of high-stakes exams. Future research may focus on a broader range of outcomes, including various measures of psychological and somatic health. Examining the long-term consequences in terms of social exclusion may also provide important information on these exams. For example, future research could include measures of employment status, such as NEET status and employment conditions, as well as measures of fertility and partnership formation, social networks, and receipt of welfare benefits. Though we showed that among those who failed, a higher proportion who were considered NEET up to 5 years after the exam compared to those who passed, a more detailed analysis of NEETs status would further our knowledge of the possible long-term consequences of high-stakes exam failure for social exclusion.

Failing a high-stakes exam has serious consequences in both the short- and long-term for adolescents. More broadly, these exams may have larger societal consequences not only through the economic costs of dropout but also through the exacerbation of social and educational inequalities. Individuals who do not complete upper secondary education experience lower lifetime earnings, lost income, and lower workforce participation, and are at an increased risk for receiving work disability benefits and having general health impairments (De Ridder et al., 2012).

The effect of high-stakes exam failure has also been shown to disproportionately impact students from low-income and minority backgrounds, which may result in the perpetuation of social and educational inequalities across generations. As a result, policymakers should be aware of the potential risks and consequences for students and society when advocating for increased use of high-stakes exams. While it may not be feasible to eliminate exams in all forms, there appears to be a trade-off between the focus on achievement and adolescent well-being. Policymakers interested in improving adolescent well-being and dropout rates, could perhaps seek alternatives to high-stakes testing or opportunities to lower the stakes. One solution may be to incorporate teacher-assessed grades and exam grades together for a final grade, rather than basing graduation on passing a single exam.

These findings also have implications for policies aimed at addressing adolescent mental health. First and foremost, school health services should be aware of the potential risk that poor school performance may place on students' mental health. Efforts should be particularly focused on the time of the year when students sit high-stakes exams or receive feedback from these, and the school health service needs to work together with the schools to consider preventive interventions. Furthermore, both primary and specialist healthcare services need to be aware of the increased vulnerability to psychological disorders that students who do not graduate are facing in the year after the exam. This may be due to both feelings of disappointment or inadequacies and lack of career and education opportunities that these young people face. Efforts should also be initiated by social services to engage non-graduates in meaningful and career-building activities after upper secondary school.

CONCLUSION

High-stakes exams can have severe consequences on the students who fail, while potentially exacerbating educational inequalities. The results show an association between failing a high-stakes exam and receiving psychological diagnoses, and therefore these exams may impact adolescents more broadly than captured in educational outcomes. Failing the exam seems to act as a sudden shock where these adolescents experience a relative increase in psychological distress compared to those who received the lowest passing grade. While the impact on mental health seems to be relatively short term, failing a high-stakes exam was associated with long-lasting impacts on dropout and enrollment in further education. These adolescents fall behind their peers not just by 1 year, but are significantly less likely to graduate from upper secondary education or enroll in tertiary education within 5 years following the exam. By using rich individual-level registry data combined with PSM methods, we have expanded the current evidence base and have examined the consequences of failing high-stakes exams for students' mental health, dropout, and enrollment in further education.

ACKNOWLEDGMENTS

This work was supported by the Research Council of Norway through its ground-breaking research funding scheme (Lost in transition? Uncovering social and health consequences of sub-optimal transitions in the education system, project No. 314562). This work was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme, project No. 262700. This study was approved by the Regional Committees for Medical and Health Research Ethics (REK) in Norway (#2018/434). We thank Dr. Magnus Nordmo for his expertise and helpful discussions during the revision of the manuscript. We also thank the discussants at the EAPS Masterclass on Families, Child Development and Well-being and the Educational Resources and Student Performance Workshop in Oslo, Norway for their insightful comments.

DATA AVAILABILITY STATEMENT

The data and materials necessary to reproduce the analyses presented here are not publicly accessible. The register data can be accessed by application to the Regional Committee for Medical and Health Research Ethics in Norway, Statistics Norway, and the Norwegian Directorate of Health. Our ethical approval does not open for storage of data on an individual level

in repositories or journals. The analytic code necessary to reproduce the analyses presented in this paper is publicly accessible, and these can be found at https://github.com/KathrynChristineBeck/DistressingTesting. The analyses presented here were not preregistered.

ORCID

Kathryn Christine Beck https://orcid.org/0000-0002-5196-4577

REFERENCES

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*, 49–74. https://doi.org/10.1037/0021-843X.87.1.49

Anderson, O. (2022). *Walking the line: Does crossing a high stakes exam threshold matter for labour market outcomes? (22-05)*. UCL Centre for Education Policy and Equalising Opportunities.

Andresen, M. E., & Løkken, S. A. (2020). *The final straw: High school dropout for marginal students*. SSRN.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*, 3083–3107. https://doi.org/10.1002/sim.3697

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*, 150–161. https://doi.org/10.1002/pst.433

Azmat, G., Calsamiglia, C., & Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, *14*, 1372–1400. https://doi.org/10.1111/jeea.12180

Borgonovi, F., Ferrara, A., S., & Maghnouj, S. (2018). *The gender gap in educational outcomes in Norway*. https://doi.org/10.1787/19939019

Caves, K., & Balestra, S. (2018). The impact of high school exit exams on graduation rates and achievement. *The Journal of Educational Research*, *111*, 186–200. https://doi.org/10.1080/00220671.2016.1226158

De Ridder, K. A., Pape, K., Johnsen, R., Westin, S., Holmen, T. L., & Bjørngaard, J. H. (2012). School dropout: A major public health challenge: A 10-year prospective study on medical and non-medical social insurance benefits in young adulthood, the young-HUNT 1 study (Norway). *Journal of Epidemiology and Community Health*, *66*, 995–1000. https://doi.org/10.1136/jech-2011-200047

Falch, T., Nyhus, O. H., & Strøm, B. (2014). Causal effects of mathematics. *Labour Economics*, *31*, 174–187. https://doi.org/10.1016/j.labeco.2014.07.016

Forgeard, M. J., Haigh, E. A., Beck, A. T., Davidson, R. J., Henn, F. A., Maier, S. F., Mayberg, H. S., & Seligman, M. E. (2011). Beyond depression: Towards a process-based approach to research, diagnosis, and treatment. *Clinical Psychology: Science and Practice*, *18*, 275–299. https://doi.org/10.1111/j.1468-2850.2011.01259.x

Goldman, D., & Smith, J. P. (2011). The increasing value of education to health. *Social Science & Medicine*, *72*, 1728–1737. https://doi.org/10.1016/j.socscimed.2011.02.047

Haavik, L., Joa, I., Hatloy, K., Stain, H., & Langeveld, J. (2017). Help seeking for mental health problems in an adolescent population: The effect of gender. *Journal of Mental Health*, *28*, 467–474. https://doi.org/10.1080/09638237.2017.1340630

Helms, J. E. (2004). Fair and valid use of educational testing in grades K-12. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 81–88). CAPS Press.

Högberg, B., & Horn, D. (2022). National high-stakes testing, gender, and school stress in Europe: A difference-in-differences analysis. *European Sociological Review*, *38*, 975–987. https://doi.org/10.1093/esr/jcac009

Högberg, B., Lindgren, J., Johansson, K., Strandh, M., & Petersen, S. (2021). Consequences of school grading systems on adolescent health: Evidence from a Swedish school reform. *Journal of Education Policy*, *36*, 84–106. https://doi.org/10.1080/02680939.2019.1686540

Högberg, B., Strandh, M., & Hagquist, C. (2020). Gender and secular trends in adolescent mental health over 24 years—The role of school-related stress. *Social Science & Medicine*, *250*, 112890. https://doi.org/10.1016/j.socscimed.2020.112890

Kumandaş, H., & Kutlu, O. (2010). High stakes testing: Does secondary education examination involve any risks? *Procedia-Social and Behavioral Sciences*, *9*, 758–764. https://doi.org/10.1016/j.sbspro.2010.12.230

Lauderdale, D. S. (2001). Education and survival: Birth cohort, period, and age effects. *Demography*, *38*, 551–561. https://doi.org/10.1353/dem.2001.0035

Leite, W. (2017). Propensity score matching. In *Practical propensity score methods using R* (pp. 87–110). SAGE Publications, Inc. https://doi.org/10.4135/9781071802854

Lorant, V., de Gelder, R., Kapadia, D., Borrell, C., Kalediene, R., Kovács, K., Leinsalu, M., Martikainen, P., Menvielle, G., & Regidor, E. (2018). Socioeconomic inequalities in suicide in Europe: The widening gap. *The British Journal of Psychiatry*, *212*, 356–361. https://doi.org/10.1192/bjp.2017.32

Machin, S., McNally, S., & Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, *190*, 104224. https://doi.org/10.1016/j.jpubeco.2020.104224

Mackenbach, J. P., Kulhánová, I., Menvielle, G., Bopp, M., Borrell, C., Costa, G., Deboosere, P., Esnaola, S., Kalediene, R., & Kovacs, K. (2015). Trends in inequalities in premature mortality: A study of 3.2 million deaths in 13 European countries. *Journal of Epidemiology and Community Health*, *69*, 207–217. https://doi.org/10.1136/jech-2014-204319

Mirowsky, J., & Ross, C. E. (2008). Education and self-rated health: Cumulative advantage and its rising importance. *Research on Aging*, *30*, 93–122. https://doi.org/10.1177/0164027507309649

Morey, J. N., Boggero, I. A., Scott, A. B., & Segerstrom, S. C. (2015). Current directions in stress and human immune function. *Current Opinion in Psychology*, *5*, 13–17. https://doi.org/10.1016/j.copsyc.2015.03.007

Nguyen, T.-L., Collins, G. S., Spence, J., Daurès, J.-P., Devereaux, P., Landais, P., & Le Manach, Y. (2017). Double-adjustment in propensity score matching analysis: Choosing a threshold for considering residual imbalance. *BMC Medical Research Methodology*, *17*, 1–8. https://doi.org/10.1186/s12874-017-0338-0

Norwegian Directorate of Health. (2020). *Fastlegestatistikk 2020* [in Norwegian]. Helsedirektoratet.no/Statistikk/Fastlegestatistikk.

Norwegian Ministry of Education and Research. (2019). *Nye sjanser—Bedre læring: Kjønnsforskjeller i skoleprestasjoner og utdanningsløp* [New chances—Better learning: Gender differences in school performance and education trajectories] [in Norwegian]. https://www.nhri.no/nou-2019-3-nye-sjanser-bedre-laering-kjonnsforskjeller-i-skoleprestasjoner-og-utdanningslop/

Ou, D. (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*, 171–186. https://doi.org/10.1016/j.econedurev.2009.06.002

Papay, J. P., Murnane, R. J., & Willett, J. B. (2010). The consequences of high school exit examinations for low-performing urban students: Evidence from Massachusetts. *Educational Evaluation and Policy Analysis*, *32*, 5–23. https://doi.org/10.3102/0162373709352530

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41

Seligman, M. E. (1972). Learned helplessness. *Annual Review of Medicine*, *23*, 407–412.

Saunes, I.S., Karanikolos, M., & Sagan, A. (2020). Norway: Health system review. *Health systems in transition*, *22*, 1–163.

Strand, B. H., Steingrímsdóttir, Ó. A., Grøholt, E.-K., Ariansen, I., Graff-Iversen, S., & Næss, Ø. (2014). Trends in educational inequalities in cause specific mortality in Norway from 1960 to 2010: A turning point for educational inequalities in cause specific mortality of Norwegian men after the millennium? *BMC Public Health*, *14*, 1–9. https://doi.org/10.1186/1471-2458-14-1208

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*. 1–28. https://doi.org/10.18637/jss.v042.i08

Wang, L. C. (2016). The effect of high-stakes testing on suicidal ideation of teenagers with reference-dependent preferences. *Journal of Population Economics*, *29*, 345–364. https://doi.org/10.1007/s00148-015-0575-7

West, P., & Sweeting, H. (2003). Fifteen, female and stressed: Changing patterns of psychological distress over time. *Journal of Child Psychology and Psychiatry*, *44*, 399–411. https://doi.org/10.1111/1469-7610.00130

World Organization of National Colleges and Academies. (2005). *ICPC-2: International Classification of Primary Care* (Rev. 2nd ed.). Oxford University Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Beck, K. C., Røhr, H. L., Reme, B.-A., & Flatø, M. (2023). Distressing testing: A propensity score analysis of high-stakes exam failure and mental health. *Child Development*, *00*, 1–19. https://doi.org/10.1111/cdev.13985