European Sociological Review, 2022, 1–13
https://doi.org/10.1093/esr/jcac009
Original Article

National High-Stakes Testing, Gender, and School Stress in Europe: A Difference-in-Differences Analysis

Björn Högberg[1,*] and Daniel Horn[2,3]

[1]Department of Social Work, Centre for Demographic and Ageing Research (CEDAR), Umeå University, SE-901 87 Umeå, Sweden, [2]Centre for Economic and Regional Studies, Institute of Economics, H-1097 Hungary and [3]Corvinus University Budapest, Institute of Economics, H-1093 Hungary

*Corresponding author. Email: bjorn.hogberg@umu.se

Submitted May 2021; revised January 2022; accepted January 2022

Abstract

In this study, we ask if high-stakes testing affects school-related stress among students and if there are gender differences in these effects. Students' results on high-stakes tests can have long-term consequences for their future educational trajectories and life chances. For girls, who tend to have higher educational aspirations and tend to gain more from higher education, the stakes involved may be even higher. The use of high-stakes testing has increased across Europe, but little is known about their consequences for stress or wellbeing. We combine macro-level data on high-stakes testing with survey data on more than 300,000 students aged 11–15 years in 31 European countries from three waves (2002, 2006, and 2010) of the Health Behaviour in School-aged Children study. With variation in high-stakes testing across countries, years, and grade levels, we use a quasi-experimental difference-in-differences design for the identification of causal effects. We find that high-stakes testing increases the risk of moving from low to high levels of self-reported school stress by 4 percentage points, or by 12 per cent relative to baseline values. This effect is somewhat larger for girls, though not significantly so. The results are robust to a range of sensitivity analyses.

Introduction

The institutionalized links between locations in the education system and locations in the overall social structure imply that the sorting of students in education systems is a crucial dimension of social stratification (Kerckhoff, 1995; Domina, Penner and Penner, 2017). Higher education confers better employment prospects, access to higher-status occupations, and greater earnings and acts as a basis for social categorization (Meyer, 1977). In order to allocate students to a stratified social structure, education systems act as gatekeepers and ration status-differentiated credentials through some form of selection procedure such as high-stakes testing (Sørensen, 1970; Domina, Penner and Penner, 2017). High-stakes tests are defined by their function to inform decisions about students' educational careers, and are used to determine grade retention, track placement, or eligibility to higher levels of the education system or certain study programmes within these (Verger, Parcerisa and Fontdevila, 2019). Because high-stakes tests are

used to ration access to more prestigious credentials, students' results on these tests can have far-reaching consequences for their future educational trajectories.

The stratification of educational credentials and opportunities is strongly gendered, and girls tend to have higher educational aspirations and to be overrepresented in academic tracks in secondary education and in tertiary education (DiPrete and Buchmann, 2013; Delaruelle, Buffel and Bracke, 2018). Girls' identity and social status is also more tightly linked to their school achievement (Mickelson, 1989; Landstedt and Gådin, 2012; Legewie and DiPrete, 2012). Because girls and women are on average disadvantaged in most social domains, they tend to be more dependent on formal educational credentials for their life chances (Ross and Mirowsky, 2006; DiPrete and Buchmann, 2013; Delaruelle, Buffel and Bracke, 2018) and may therefore face greater pressure to perform well on high-stakes tests in order to secure access to such credentials.

While a substantial amount of research has investigated the consequences of educational stratification or high-stakes testing on academic or employment outcomes (Scheeren, van de Werfhorst and Bol, 2018; Phelps, 2019; van Hek *et al.*, 2019), research on non-academic outcomes such as the well-being of students is scarce. In this study, we focus on school-related stress. Due to the long-term consequences for future educational trajectories, high-stakes testing may generate stress and, by extension, poor mental health (Pearlin *et al.*, 1981). Though large-scale and systematic evidence is lacking, ethnographic studies show that high-stakes testing looms over the lives of adolescents, to the extent that their 'whole life' depends on them performing well (Banks and Smyth, 2015). Accordingly, school is consistently ranked among the greatest stressors for adolescents and girls in particular (Byrne, Davenport and Mazanov, 2007). Moreover, both high-stakes testing and self-reported school stress—understood as subjective stress responses to demands that emanate from school—have increased in recent decades across Europe (Verger, Parcerisa and Fontdevila, 2019; Löfstedt *et al.*, 2020).

Against this background, the present study investigates the effects of national high-stakes testing on self-reported school stress among adolescents, with a specific focus on gender inequalities. To this end, we use repeated cross-sectional survey data on more than 300,000 students (aged 11–15 years) in 31 European countries from 2002 to 2010. We use a quasi-experimental difference-in-differences design to identify causal effects, with variation in high-stakes testing across countries, years, and grade levels.

This study makes two key contributions to existing research on educational stratification, high-stakes testing, and school stress. First, we extend the literature on the consequences of educational stratification and high-stakes testing by studying novel non-academic outcomes. Recent decades have seen a growth in the use of high-stakes testing as a means to sort students in the education system, with the yearly number of national high-stakes tests conducted across OECD countries doubling from 1995 to 2014 (Verger, Parcerisa and Fontdevila, 2019). Non-academic outcomes are increasingly being recognized as important dimensions for education systems (Montt and Borgonovi, 2018), and knowledge regarding the effects of testing policies on well-being is relevant for policymakers and researchers alike.

Second, we deepen the understanding of how and why gender inequalities in school stress emerge during early adolescence. Our quasi-experimental difference-in-differences design improves on causal inferences compared to the often cross-sectional literature on sources of gender inequalities in school stress (Byrne, Davenport and Mazanov, 2007; Sonmark *et al.*, 2016). We also provide policy-relevant knowledge regarding how such gender inequalities may, or may not be, addressed. If girls are harmed more by high-stakes testing, policymakers have a tangible tool to promote gender equality; if not, abolishing high-stakes testing will probably have little effect.

Background and Previous Research

School Stress

The available data limit the conceptualization of stress used in this study (see below). However, in a general sense, stress is understood as a subjective response to demands that are perceived as unmanageable and threatening (cf. Pearlin *et al.*, 1981). School stress is understood as stress responses to demands that emanate from the school, while high-stakes tests are understood as stressors in the school context. While stress in school may have some positive consequences, for instance, by pushing students to study, research shows that school stress is strongly correlated with psychosomatic symptoms and poor psychological well-being (Sonmark *et al.*, 2016; Högberg *et al.*, 2019) and anxiety and depression (Byrne, Davenport and Mazanov, 2007). Girls report more stress, and school stress is an important factor behind the gender gap in mental health—i.e. that girls report more mental health problems than boys—that

opens up during adolescence (Högberg, Strandh and Hagquist, 2020).

Previous Research

A few studies from the United States have investigated the effects on stress or similar outcomes of testing policies linked to school accountability laws. Comparing stress levels over the school year, Heissel *et al.* (2021) and Segool *et al.* (2013) found stress to be higher in periods when such tests are conducted, though neither included a comparison group that was not exposed to testing. Whitney and Candelaria (2017) investigated the staggered adoption of testing policies across US states and found weak positive effects on self-reported anxiety. Results on gender inequalities are mixed, and Heissel *et al.* (2021) found weaker, but Segool *et al.* (2013) found stronger effects for girls, while Whitney and Candelaria (2017) found no gender differences in effects. However, tests linked to school accountability laws are primarily high-stakes for schools but not for students because their results typically have no or weak consequences for their future educational careers. To the best of our knowledge, no study has investigated the effects of tests that are high-stakes for the students themselves and that predominate in the more stratified European education systems.

Theoretical Framework

The effects of high-stakes testing on stress can be understood from the perspective of the dual role that education systems play in shaping social stratification processes (Meyer, 1977; Kerckhoff, 1995; Domina, Penner and Penner, 2017). First, all education systems are hierarchically ordered in various ways. Grades or educational stages are defined as progressions from preceding ones, certain study programmes or tracks are defined as more advanced than others, and marks and test results are used to stratify students based on achievement. From the perspective of theories of categorical inequality (Domina, Penner and Penner, 2017), this hierarchical structure, and the credentials used to demarcate its boundaries, can be conceptualized as bringing forth corresponding social categories, such as grade repeaters, dropouts, vocational students, and high-performing students. Because the categories are based on an officially sanctioned hierarchy, they imply a legitimate status distinction between students sorted into them (Meyer, 1977).

Second, the hierarchical structure of the education system is interlocked with other systems of stratification. Because educational credentials are used to allocate individuals to locations in the labour market and other stratified systems, the social categories created by the education system are translated into positions in a hierarchy of social prospects (Kerckhoff, 1995; Domina, Penner and Penner, 2017). Moreover, the sequential and path-dependent nature of the education system implies that what happens at key branching points at lower educational stages also has consequences by limiting eligibility at higher stages, thus constraining social prospects in adulthood (Breen and Jonsson, 2000; Härkönen and Sirniö, 2020).

In many education systems, high-stakes testing is integral to this dual stratification process. First, by awarding certificates and determining eligibility to higher stages, programmes, or tracks, they sort students into salient social categories. Second, because high-stakes testing is used as a selection instrument at key branching points, it regulates educational trajectories also in the long term, implying that these social categories have implications beyond the education system itself (e.g. access to high-status occupations). There are reasons to expect that these two processes may combine to generate stress among students.

Concerning the first process (sorting into social categories), qualitative studies show that being categorized as a high-achieving student is salient for students' identity, social status, and mental health (Mickelson, 1989; Reay and Wiliam, 1999; Landstedt and Gådin, 2012). Because tests make achievement more explicit, and high-stakes tests, in addition, link this achievement to access to formal credentials, they increase the salience of the associated social categories. Moreover, a high workload, with time-consuming homework and test preparation, is considered among the most stressful aspects of school (Byrne, Davenport and Mazanov, 2007). Due to their perceived importance, high-stakes tests are likely to amplify this workload (Lee and Larson, 2000; Banks and Smyth, 2015).

Concerning the second process (translation of social categories into social prospects), the important aspect here is that students also at lower stages of the education system are highly aware of the importance of educational credentials for their educational and employment prospects. Accordingly, several studies have found that students in the focal age category for this study view their achievement on high-stakes tests as crucial for their life chances and that this is one of the main reasons why these tests are experienced as more stressful than other types of assessments (Reay and Wiliam, 1999; Denscombe, 2000; Lee and Larson, 2000; Banks and Smyth, 2015; Högberg *et al.*, 2019).

Gender Differences in School Stress

Gender inequalities in school stress can also be understood from the perspective of the two school-related

hierarchies discussed above. As for the first (sorting into social categories), gender roles differentially shape how girls and boys relate to school-based social categories (Domina, Penner and Penner, 2017). Being a 'good' and high-achieving student is typically viewed as feminine and thus more compatible with gender roles available to girls. In contrast, identities such as being an athlete or a rebel are viewed as masculine and valued higher for boys (Landstedt and Gådin, 2012; Legewie and Di Prete, 2012). Moreover, girls tend to be more sensitive to social approval and extrinsic rewards in school (Mickelson, 1989), including officially sanctioned credentials such as those granted through high-stakes tests. Girls also tend to have higher educational aspirations and often expect to enrol in academic tracks or higher education (DiPrete and Buchmann, 2013). To satisfy such expectations, they typically need good test results if the selection is regulated by high-stakes testing.

Regarding the second process (translation of social categories into social prospects), resource substitution theory states that a given resource provides relatively greater gains when few alternative resources can be used to reach the same outcome (Ross and Mirowsky, 2006; Delaruelle, Buffel and Bracke, 2018). Girls and women are on average disadvantaged in most social domains, and girls may suffer from a relative lack of other resources (e.g. social capital) when entering the labour market. They may then be more dependent on formal educational credentials and the test results needed to obtain them. Accordingly, women tend to have higher returns from education such that women gain relatively more than men in terms of earnings, marriage, and protection against poverty (Pekkarinen, 2012; DiPrete and Buchmann, 2013). High-stakes tests may be therefore 'higher-stakes' tests for girls. Moreover, studies show that girls' educational achievement and attainment are impeded by earlier selection into educational tracks (Scheeren, van de Werfhorst and Bol, 2018; van Hek *et al.*, 2019). If high-stakes tests are used to guide this selection, girls may then experience them as more stressful.

Data and Methods

National testing is by definition shared by all students in an education system and, therefore, is difficult to study in a single country at a single point in time. With harmonized cross-country data spanning over years and grades, we can use variations in testing policies across countries, grades, and time.

Individual-Level Data

We use individual-level survey data from the Health Behaviour in School-aged Children (HBSC) study. HBSC is a repeated cross-sectional and standardized survey of students aged 11–15 years conducted every 4 years in collaboration with the World Health Organization. HBSC is among the most comprehensive international adolescent health surveys and has been extensively used in comparative research on school stress (e.g. Löfstedt *et al.*, 2020). HBSC collects representative data on three age groups—11.5, 13.5, and 15.5 years on average—corresponding to three country-specific grades with a typical sample size of around 5,000 students per country and survey (Roberts *et al.*, 2009). We use the maximum number of countries and survey years (2002, 2006, 2010) for which we have matching data on national testing policies (see below), resulting in a dataset with more than 300,000 students in 31 European countries.

High-Stakes Testing

We use data from Eurydice (2009) to identify national high-stakes testing, defined as tests where the outcome of the test has appreciable consequences for students' educational trajectories. Eurydice is an European Union network with the task of providing comparable information on European education systems. Eurydice data on national testing have been validated in previous comparative studies (Braga, Checchi and Meschi, 2013). Because these data are not available after 2010, we cannot use the most recent HBSC surveys. Eurydice covers compulsory and nationally defined (regionally in the case of Belgium and Great Britain) tests standardized by top-level education authorities.

Eurydice data have two advantages given the aim of this study. First, they distinguish between different types of national tests depending on their purpose. Specifically, they provide information on tests used for 'making decisions about the school career of pupils,' which are defined as tests used to determine grade retention, track placement, or eligibility to higher levels of the education system. We use this as treatment indicator because the direct consequences for students' educational trajectories imply that they have high-stakes. Second, Eurydice data provide information on the year in which the test was first implemented and the grade in which students take the test, thus enabling us to use variation across time and grades in addition to variation across countries for estimation. It should be noted that we only have individual data on three grades and cannot identify the effects of tests given in other grades.

School Stress

We conceptualize school stress as perceptions of excessive demands related to school. The HBSC data contain one item—'*How pressured do you feel by the schoolwork you have to do?*'—that measures the global feeling of being pressured by schoolwork. We dichotomize this, with response options 'Not at all' and 'A little' coded 0 and 'Some' and 'A lot' coded 1. Supplementary Figures S1 and S2 show the distribution of the variable in all included countries. While a more detailed measure directly focusing on the role of testing would have been desirable, the item has some advantages for this study. Along with two other items—'I find the schoolwork difficult' and 'The schoolwork makes me tired', which are not available in all countries—the item has been included in validated subscales measuring school stress (Löfstedt *et al.*, 2020). Among these three items, the one used here is the strongest predictor of health (Sonmark *et al.*, 2016) and has been used *in lieu* of the full subscale to track levels and correlates of school stress across countries (Löfstedt *et al.*, 2020). For these reasons, we believe that the item may serve as a valid proxy for school stress.

Covariates at the Individual Level

A key individual-level variable is the grade of the student. HBSC collects data on three groups aged on average 11.5, 13.5, and 15.5 years (henceforth 11, 13, and 15 years), which correspond to the typical grades for students in these age groups in the respective countries. In some countries, students in the relevant age group may be spread over more than one grade due to grade retention. However, most participating countries in HBSC only draw samples from the age-typical grades, and only about 1 per cent of the sample differ by more than 1 year from the expected average age (11.5, 13.5, and 15.5 years). While this indicates that grade repeaters may be undersampled in HBSC, potential undersampling likely does not bias the results because the share of students repeating a grade in secondary school (data from Eurydice (2011)) is similar in countries with or without high-stakes testing (6.4 per cent vs. 7.7 per cent).

Our identification strategy primarily relies on using fixed effects for countries, grades, and years (see below). To adjust for potential remaining sources of confounding, we want to adjust for factors associated with both testing and stress but that are not themselves consequences of testing, net of the fixed effects. Note that we cannot account for pre-treatment values of potential individual-level confounders with cross-sectional data at the individual level. To account for compositional differences across student populations, we control for age, gender, whether the student lives with both parents, the consumption level of the household, and parental occupation. These 'demographic' controls are unlikely to be affected by testing policies. We also control for a range of individual characteristics that may at the group level (country, grade, or year) be associated with testing policies, and, at the individual level, be associated with stress. Because existing research on correlates or determinants of high-stakes testing policies is rudimentary, the choice of these 'additional' student-level controls is primarily based on theoretical considerations. For instance, policymakers may use high-stakes testing as a disciplinary tool in countries, grades, or years when disruptive behaviours are common. We, therefore, adjust for the frequency of binge drinking, bullying, physical fighting, and physical injury. It is also possible that high-stakes testing is more common in countries, grades, or years where determinants of stress, which are themselves not causally related to testing, are more common. We, therefore, adjust for the quality of child–parent relationships, the student's opinion of his/her body, and frequency of physical activity. Unlike the demographic controls, the latter set may potentially be caused by testing, and it is not evident whether they should be regarded as confounders or mediators of the relationship between testing and stress. We therefore present results with and without these additional controls included.

Covariates at the Country Level

While research on macro-level determinants of high-stakes testing policies or school stress is scarce, we adjust for a set of time-varying country-level indicators that have been shown to be related to adolescent well-being more broadly (e.g. Elgar *et al.*, 2015), namely gross domestic product per capita, country-level economic inequality as measured by the GINI index, youth unemployment rate, and the tertiary attainment rate among young adults (aged 25–34 years) (data from Eurostat (2020a,b,c) and OECD (2020)).

The exact measurement (survey questions and response options) and descriptive statistics for all individual and country-level variables are presented in Supplementary Tables S1 and S2.

Analytical Strategy

We use a difference-in-differences design for estimation. Because the data vary across countries, years, and student grades, we use multiple comparison groups—grades and years—within countries, resulting in a triple-

differences design. As is evident from Supplementary Table S3, there is variation in high-stakes testing across all three dimensions. A total of 16 out of the 31 included countries have high-stakes testing in some grade and at some time during the study period (2002–2010). In 11 of these, high-stakes tests are taken at age 15, in 2 at age 13, and in 3 at age 11. Six countries (Belgium (Walloon region), Germany, Iceland, Italy, Norway, and Romania) introduced or abolished high-stakes testing between 2002 and 2010. We utilize these sources of variation combined in our main estimation and separately in the supplementary analyses.

Based on the three sources of variation combined, we estimate a three-way fixed effects regression model of the following form:

$$\gamma_{igcy} = \beta_0 + \beta_1 C_c + \beta_2 G_g + \beta_3 Y_y + \beta_4 T_{gcy} + \beta_5 X_{igcy} + \beta_6 Z_{cy} + \varepsilon_{icyg} \quad (1)$$

where i stands for the individual student, g for grade, c for country, and y for survey year.

A full set of dummy variables is included for each of the three sources of variation. Dummy variables for countries are denoted by C_c and capture time and grade-invariant differences across countries. Dummy variables for grades are denoted by G_g and capture time and country-invariant differences across grades. Dummy variables for survey years are denoted by Y_y and capture temporal changes that are invariant across countries and grades. The focal explanatory variable—high-stakes testing—is denoted by T_{gcy} and is coded 1 for students in countries, grades, and years with high-stakes testing and 0 otherwise. Thus, β_4 gives the effect of high-stakes testing and is identified based on variation in testing either across grades or time within countries. To account for possible confounding that may also vary across grades or time within countries, we include X_{igcy}, which is a vector of individual-level covariates, and Z_{cy}, which is a vector of time-varying country-level covariates, as described previously. ε_{icyg} is an individual-specific error term.

We also extend Eq. 1 by including a full set of two-way interactions between the dummies for country, survey year, and grade in a fully saturated triple-difference model:

$$\gamma_{igcy} = \beta_0 + \beta_1 C_c + \beta_2 G_g + \beta_3 Y_y + \beta_4 T_{gcy} + \beta_5 C_c{*}Y_y + \beta_6 G_g{*}Y_y + \beta_7 C_c{*}G_g + \beta_8 X_{igcy} + \varepsilon_{igcy} \quad (2)$$

This allows us to control non-parametrically for all sources of confounding that do not vary across all three dimensions. Specifically, estimation is based on taking the change over time in stress for students in a country and grade that implement testing in that grade in a specific year, net of the corresponding change over time for students in the same grade in other countries (G × Y), the change for students in other grades in the same country (C × Y), and the difference between students in the same grade and country but in different years (G × C). We have 258 distinct country-grade-year groups with data on testing, 188 of which are subsumed by the fixed effects in the fully saturated model. Note that the country-level covariates in Z_{cy} (in Eq. 1) drop out from Eq. 2 because they do not vary across grades. To study gender differences, we estimate the equation separately for girls and boys.

We also investigate the country-by-grade and country-by-year variation separately to see if the different sources of variation in testing generate similar results. Country-by-grade variation is investigated by limiting the sample to countries and years in which testing is conducted in at least one grade. Thus, if a country introduced testing in 2009, we only use the 2010 survey (but all grades) for that country, ensuring that the sample contains no variation in testing across time. Country-by-year variation is investigated by limiting the sample to countries and grades in which testing is either introduced or abolished over the study period. Thus, if a country introduced testing at age 15 in 2009, we only include this grade level (but all years) for that country, ensuring that the sample contains no grade-level variation in testing. We then estimate a modified version of Eq. 1 on these subsamples, where we drop the 'y' subscripts and year fixed effects ('Y') when only utilizing variation across grades and drop the 'g' subscripts and the grade fixed effects ('G') when only utilizing variation across time.

We use a linear least square dummy variable estimator with cluster robust standard errors to account for the dependence of individual observations within clusters. We follow the design-based approach to clustering suggested by Abadie *et al.* (2017), and cluster at the level of treatment assignment, which is grades within countries (93 possible combinations with 31 countries and three grades). We use wild cluster bootstrap for inference because this is more conservative and performs better in finite samples than default cluster robust standard errors. Simulations show that wild cluster bootstrap produces reliable inference with as few as 20 clusters, well below the 93 clusters used in this study (Cameron and Miller, 2015). Cluster robust standard errors are also heteroskedasticity-consistent, which is important because we use linear regression with a binary outcome variable.

The primary assumption required for drawing causal conclusions from a difference-in-differences analysis is the parallel trends assumption. This implies that the differences in outcomes between treatment and control groups would be constant in the absence of treatment. Because the assumption refers to counterfactual outcomes that cannot be observed, it cannot be tested directly. A standard way to probe if the assumption is reasonable is to compare pre-treatment trends in the treatment and control groups. We only have three grade levels and time periods, and most countries have introduced high-stakes tests previous to our earliest available individual-level data in 2002. Given these restrictions, comparisons of pre-treatment trends are hardly informative. Note, however, that the fully saturated triple-difference model described in Eq. 2 makes the assumption less demanding (Wooldridge, 2010). In a model that only uses variation over time, the parallel trends assumption implies that, if high-stakes testing was not introduced, the difference in stress between testing and non-testing countries would be constant. In such a case, unobserved country-specific time shocks could generate bias. When we add variation across grades and estimate a triple-difference model, the parallel trends assumption instead implies that, if high-stakes testing was not introduced in a certain grade, the difference in stress across grades would otherwise have evolved in the same way in testing and non-testing countries; or similarly, the difference in stress over time would otherwise have been the same across grades in testing and non-testing countries. Thus, the triple-difference model allows for unobserved country-specific time shocks as long as they affect all grades similarly, and for unobserved differences across grades as long as they are constant over time.

An additional assumption is that the treatment does not have spillover effects on students in the control groups. This is related to the stable unit treatment value assumption, that is, that treatment of one unit does not affect the outcomes of non-treated units. Such spillover effects would be present if students in grades with no testing are affected by the stress experienced by students in grades with testing, or if the anticipation of tests in upper grades causes stress. The most likely scenario would be that students in grades with no testing experience higher stress due to upcoming tests, which would bias the estimate of the effect of testing downwards. Another potential threat is reverse causality, in that high-stakes testing is introduced when reported school stress is high. Available data suggest that high-stakes testing in Europe is primarily used to sort students or to ensure that students meet learning goals (Eurydice, 2009; Verger, Parcerisa and Fontdevila, 2019). We find it unlikely that tests would be introduced because stress is deemed to be too low.

Results

We present the main results in Table 1. The table contains nine models, with stepwise inclusion of fixed effects and country and student-level covariates. Our baseline model 1 includes all cases and the treatment variable as the sole independent variable, while model 2 restricts the sample to complete cases to enable comparison with models with additional covariates. Testing is associated with around an 8 percentage point increase in stress, corresponding to a 25 per cent relative increase compared to baseline values. Model 3 adds fixed effects for countries, years, and grades (Eq. 1, but without covariates), which reduces the estimated treatment effect to 4.9 percentage points (14 per cent) but simultaneously makes the estimate more precise. Models 4–6 include demographic covariates, macro-level covariates, and the additional student-level covariates in a stepwise manner. The estimated effect is substantively unchanged through models 4–6. Considering recent findings that including covariates in two- or three-way fixed effects models introduce additional assumptions regarding homogeneous (in covariates) treatment effects and no covariate-specific trends (Callaway and Sant'Anna, 2020), it is reassuring that the results are very similar regardless of whether covariates are included or not.

The fully-saturated model 7 adds a complete set of two-way interactions between country, grade, and year dummies, thereby estimating Eq. 2. Again, the estimated effect is substantively unchanged, showing that testing increases stress by 4.1 percentage points. Using model 7, the predicted risk of stress without high-stakes testing is around 34 per cent, compared to 38 per cent with high-stakes testing, implying a 12 per cent relative increase in stress due to high-stakes testing.

All in all, we find that high-stakes testing has a statistically significant and non-negligible effect on school stress. This effect is an average effect for both genders, but based on theoretical considerations, we expect the effect for girls to be stronger. Models 8 and 9, therefore, estimate Eq. 2 separately for girls and boys. The estimated effect for girls is 4.9 percentage points (14 per cent) compared to 3.2 percentage points (9 per cent) for boys. Thus, the estimated effect is more than 50 per cent larger for girls. However, *post-hoc* calculations show that the difference between girls and boys is not statistically significant.

Because our identification strategy relies on variation over time and across grades for estimation, we re-

Table 1. Linear regression models with school stress as the outcome

	m1	m2	m3	m4	m5	m6	m7	m8 Girls	m9 Boys
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
High-stakes testing	0.079	0.085	0.049**	0.048**	0.048**	0.044**	0.041***	0.049***	0.032**
	(0.041)	(0.045)	(0.014)	(0.014)	(0.014)	(0.013)	(0.010)	(0.013)	(0.010)
Intercept	0.337	0.338	0.088	0.043	0.234	0.154	0.321	0.489	0.179
Country, year, grade fixed effects			Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-way interactions: country, year, grade							Yes	Yes	Yes
Demographic covariates				Yes	Yes	Yes	Yes	Yes	Yes
Country covariates					Yes	Yes			
Additional student covariates						Yes	Yes	Yes	Yes
Sample	Full	Complete cases	Complete cases	Complete cases	Complete cases	Complete cases	Complete cases	Only girls	Only boys
N individual level	416,003	317,834	317,834	317,834	317,834	317,834	317,834	164,522	153,312
N country-grade level	93	93	93	93	93	93	93	93	93

Notes: Cluster robust standard errors in parentheses. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006, and 2010.

***$P < 0.001$,

**$P < 0.01$,

*$P < 0.05$.

b, regression coefficient; se, standard error.

estimated the models using these sources of variation separately. Estimates using country-by-grade variation are very similar to those in Table 1. Testing significantly increases stress by 4.0 percentage points on average for both genders and by 4.5 and 3.5 percentage points for girls and boys separately (Supplementary Table S4). Effect sizes are smaller when based on variation across years, namely 2.3 percentage points on average for both genders and 1.5 and 3.2 percentage points for girls and boys separately (Supplementary Table S5), with neither estimate being significant. We also see that the standard errors are around twice as large.

Supplementary and Sensitivity Analyses

Stress may be experienced as positive by providing motivation (i.e. eustress), in which case increased stress due to testing would not necessarily be negative. In Supplementary Table S6, we show how our stress indicator relates to three common indicators of adolescent well-being: life satisfaction, self-rated health, and a screening instrument for psychosomatic symptoms. Since stress and the well-being indicators are measured simultaneously, these estimates cannot be interpreted as causal, but they indicate whether our stress indicator primarily captures positive or negative aspects of stress. The results show that moving from the lowest to the highest response option of the stress indicator is associated with 0.498–0.663 lower standard deviations in life satisfaction and self-rated health and 0.966 higher standard deviations in psychosomatic symptoms. Moreover, Supplementary Table S7 shows that testing reduces the risk of the lowest response option ('Not at all') by around 3 percentage points and increases the risk of the two highest response options ('Some' and 'A lot') by 1.6 and 1.7 percentages points, respectively. In relative terms, the effect is strongest for the option 'A lot'.

Another way to investigate whether testing primarily captures negative aspects of stress is shown in Supplementary Table S8, where we re-estimate Eq. 2 but replace stress with two other indicators of school-related well-being: school satisfaction and school climate. Testing has a weak but insignificant negative effect on school satisfaction and a moderate negative effect on the school climate, but this is not significant for girls.

We have also tested the credibility of the key assumptions of the research design in several ways. First, we estimated a set of placebo tests, and we re-estimated Eq. 2 with lead values for the treatment indicator. Specifically, we manipulated the high-stakes testing indicator so that the grade at which the test is conducted is moved two grades 'upwards', meaning that the indicator is coded 1 for 13-year-old students in countries with tests for 11-year-old students and so on. The results (columns 1–3 in Supplementary Table S9) show that this manipulated variable does not affect stress. This also indicate that spillover effects are not a major issue, that is, there are no effects of anticipation of tests in upper grades. Second, we replaced the high-stakes testing indicator with an equivalent indicator for national testing for other purposes that are not high-stakes for students. Eurydice (2009) provides data on tests used for 'monitoring schools and/or the education system' and tests used for 'identifying individual learning needs'. The first of these includes tests used to evaluate schools' performance, which may be high-stakes for schools. The second includes formative tests used to identify whether students reach stated learning goals and that are low-stakes for both students and schools. The results (columns 4–6 in Supplementary Table S9) show that these low-stakes tests have no effect on stress. We have also re-estimated the models while excluding countries with high rates of grade retention (exceeding 20 per cent in lower secondary school) because the correspondence between age group and grade may be weaker when grade retention is common. Excluding these countries does not affect the results in a substantial way (columns 1–3 in Supplementary Table S10). The results are also similar if we include country-specific linear time trends in models estimating Eq. 1 (columns 4–6 in Supplementary Table S10) and if we use the stress variable as a continuous variable (columns 7–9 in Supplementary Table S10).

Because high-stakes testing is integral to selection processes in many stratified education systems, we have also investigated whether the effects of testing are confounded by overall levels of educational inequality using achievement data from Angrist *et al.* (2021). Specifically, we control for average and standard deviation in achievement scores (to adjust for overall inequality) and for the achievement scores for boys and girls separately (to adjust for gender inequality). Supplementary Table S11 shows that the average effects are not much affected by adjusting for achievement scores, and the reduced effect size in column 4 is almost solely due to the loss of about half of the observations when conditioning on availability of achievement data (compare column 3 with 4). The gender-specific effects are more variable, but overall the gender difference is larger when adjusting for achievement scores (Supplementary Table S12).

Recent methodological literature stresses that two- or three-way fixed effects models may be biased due to heterogeneous treatment effects. Supplementary Table

S13 shows that the estimates are very similar when we use recently proposed imputation-based methods to account for this (Borusyak, Jaravel and Speiss, 2021).

Discussion and Conclusions

This study aimed to investigate the effects of national high-stakes testing on school stress among adolescent students, with a specific focus on gender differences. The results showed that high-stakes testing increased self-reported school-related stress by around 4 percentage points or around 12 per cent. This average effect was slightly but not significantly stronger for girls. The results were robust to a range of sensitivity analyses, including placebo tests using lead values or indicators of low-stakes tests, excluding countries with potentially lower quality data, and accounting for heterogeneous treatment effects. We have also ruled out reverse causality and spillover effects.

Positive effects of high-stakes testing on stress are consistent with qualitative European studies showing that students report high-stakes testing to be among their most stressful experiences in school (Banks and Smyth, 2015) and also with some quantitative American studies (Segool *et al.*, 2013; Heissel *et al.*, 2021). These results are in line with theoretical predictions. The fact that education systems are both internally stratified—by categorizing students in an officially sanctioned hierarchy—and that this stratification is interlocked with broader stratification systems entails that selection processes that sort students in the education system are perceived as stressful and threatening by students.

The small and non-significant gender difference in the effect is consistent with Whitney and Candelaria (2017) but contradicts predictions based on theories of gender roles in schools and resource substitution theory (Mickelson, 1989; Ross and Mirowsky, 2006). One way to reconcile our findings with these theoretical perspectives is to note that the point estimate was more than 50 per cent larger for girls but that this study is underpowered to detect significant heterogeneous effects. Another explanation is that girls tend to perform better in school and therefore may feel that they have greater chances to succeed on high-stakes tests.

The estimated effects are non-negligible and can be regarded as problematic. We found that testing increased the risk of stress by around 4 percentage points, or 12 per cent compared to baseline values. This is similar to the average gender gap in stress. Considering that girls consistently tend to report substantially more school stress (Banks and Smyth, 2015; Högberg Strandh and Hagquist, 2020; Löfstedt *et al.*, 2020), this is certainly not trivial. We also found that our stress indicator is strongly negatively correlated with health and well-being and that testing also leads to a poorer school climate, although not to lower school satisfaction. The strong effects on stress and null findings on school satisfaction may seem contradictory. A possible explanation is that the effects of high-stakes testing vary depending on student engagement. Engaged students may feel stressed but also motivated and self-confident, while disengaged students may dissociate from the situation and feel alienated. Note, however, that psychological studies show that a substantial share of engaged students tend to report harmful stress, exhaustion, and mental health problems, despite valuing school highly, while an equally substantial share of disengaged students report low stress and good mental health, despite finding school meaningless (Tuominen-Soini and Salmela-Aro, 2014). Thus, although the stress generated by high-stakes testing does not seem to affect students' school satisfaction, it may nevertheless be damaging for their mental health more broadly.

We used variations in testing across countries, grades, and years for estimation. More fine-grained analyses showed that estimates based on country-by-grade variation in testing were larger than estimates based on country-by-year variation. This could be because most country-by-grade variation is due to countries using high-stakes testing at age 15, presumably to regulate the transition to upper secondary school, while much of the country-by-year variation is due to tests introduced in earlier grades. Older students may be more aware of the implications of testing, and the transition to upper secondary school may be considered to be particularly consequential and stressful. However, we believe that the difference is most likely due to methodological reasons. First, the country-by-year variation is smaller, which makes the estimates less precise and more difficult to compare. Second, estimates based on only one source of variation are more likely to be biased due to unobserved heterogeneity compared to estimates that combines both sources of variation in a triple-difference model (Wooldridge, 2010). We therefore put more faith in the estimates from the triple-difference model.

Limitations

The results of this study should be interpreted in light of its limitations. Eurydice only covers national tests, and if national tests are correlated with other (e.g. regional) tests that are experienced as high-stakes by students, this may lead to bias. Related to this, there may be a

temporal mismatch between the data collection date in HBSC and the date of the high-stakes tests (both of which vary across countries). Potential temporal mismatches will most likely make the estimates noisier and lead to attenuation bias because students will be less stressed when tests are temporally distant.

We could only measure stress with a single self-reported indicator. While this indicator has shown desirable properties in previous studies (Sonmark *et al.*, 2016), a comprehensive set of items, or biomarkers such as cortisol levels (Heissel *et al.*, 2021), would have been more desirable. Moreover, the indicator used may empirically capture both a permanent state of chronic stress and a temporary period of pressure. Chronic stress is more harmful to health (Pearlin *et al.*, 1981), while temporary spells of pressure may improve achievement without negative effects on health. However, the fact that the indicator is strongly correlated with psychosomatic symptoms suggests that the stress seems to be chronic for at least some students. A related issue is whether instantaneous and long-term effects of testing are equivalent. The first cohort exposed to a newly introduced test may be more stressed than subsequent cohorts, for whom the test is viewed as less extraordinary (cf. Högberg *et al.*, 2019). If this is so, we should observe larger effects using only variation across time (which captures the introduction of new tests for a specific grade) than variation across grades (which captures variation in exposure to 'old' tests for different grades), but this is not the case here (see Supplementary Tables S4 and S5). Alternatively, the stress may subside after students have taken the test. This would be in line with the underlying theory because part of the stress associated with high-stakes testing emanates from the intense test preparations (Banks and Smyth, 2015).

Implications

We have shown that national high-stakes testing increases school stress, and thus possibly in extension mental health problems (Pearlin *et al.*, 1981). If student well-being is a prioritized goal for education policy, these findings have implications for the optimal design of education systems. To the extent that high-stakes testing increases achievement (Phelps, 2019), policymakers and schools may face tradeoffs between achievement and well-being (Montt and Borgonovi, 2018). Likewise, because high-stakes testing is used to sort students in stratified education systems, the findings in this study have implications for policymakers interested in how such stratification shapes the broader experiences of students in school. Policymakers that value student well-being would be advised to consider alternatives to high-stakes testing or ways to lessen the stress caused by testing. In cases where testing is used for secondary level placements, a less rigid sequential structure in the education system could make the tests less high-stakes and less stressful. This could involve making it easier to change between educational programmes or introducing second chance opportunities for students who fail at specific critical junctures.

Supplementary Data

Supplementary data are available at *ESR* online.

Funding

The work on this study (planning, analysis, and writing) was funded by the Swedish Research Council [grant number 2018-03870_3].

References

Abadie, A. *et al.* (2017). *When Should You Adjust Standard Errors for Clustering?* NBER Working Paper No. w24003. National Bureau of Economic Research.

Angrist, N. *et al.* (2021). Measuring human capital using global learning data. *Nature*, **592**, 403–408.

Banks, J. and Smyth, E. (2015). 'Your whole life depends on it': academic stress and high-stakes testing in Ireland. *Journal of Youth Studies*, **18**, 598–616.

Borusyak, K., Jaravel, X. and Speiss, J. (2021) *Revisiting Event Study Designs: Robust and Efficient Estimation*. Working paper.

Braga, M., Checchi, M. and Meschi, E. (2013). Educational policies in a long-run perspective. *Economic Policy*, **28**, 45–100.

Breen, R. and Jonsson, J. O. (2000). Analyzing educational careers: a multinomial transition model. *American Sociological Review*, **65**, 754–772.

Byrne, D. G., Davenport, S. C. and Mazanov, J. (2007). Profiles of adolescent stress: the development of the adolescent stress questionnaire (ASQ). *Journal of Adolescence*, **30**, 393–416.

Callaway, B. and Sant'Anna, P. (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, **225**, 200–230.

Cameron, C. A. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50**, 317–372.

Delaruelle, K., Buffel, V. and Bracke, P. (2018). The reversal of the gender gap in education: what does it mean for gender differences in the relationship between education and health. *European Sociological Review*, **34**, 629–644.

Denscombe, M. (2000). Social conditions for stress: young people's experience of doing GCSEs. *British Educational Research Journal*, **26**, 359–374.

DiPrete, T. A. and Buchmann, C. (2013). *Rise of Women: The Growing Gender Gap in Education and What It Means for American Schools*. New York: Russell Sage Foundation.

Domina, D., Penner, A. and Penner, E. (2017). Categorical inequality: schools as sorting machines. *Annual Review of Sociology*, **43**, 311–330.

Elgar, F. J. *et al.* (2015). Socioeconomic inequalities in adolescent health: a time-series analysis of 34 countries participating in the Health Behaviour in School-aged Children study. *The Lancet*, **385**, 2088–2095.

Eurostat (2020a). Gini coefficient of equivalised disposable income – EU-SILC survey. [accessed 21 October 2020]. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di12

Eurostat (2020b). Unemployment by sex and age, Less than 25 years, percentage of active population. [accessed 21 October 2020]. https://ec.europa.eu/eurostat/databrowser/view/une_rt_a_h/default/table?lang=en

Eurostat (2020c). Population by educational attainment level, sex and age (%) – main indicators. [accessed 21 October 2020]. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=edat_lfse_03&lang=en

Eurydice. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Brussels: Education, Audiovisual and Culture Executive Agency.

Eurydice. (2011). *Grade Retention during Compulsory Education in Europe: Regulations and Statistics*. Brussels: Education, Audiovisual and Culture Executive Agency.

Härkönen, J. and Sirniö, O. (2020). Educational transitions and educational inequality: a multiple pathways sequential logit model analysis of Finnish birth cohorts 1960–1985. *European Sociological Review*, **36**, 700–719.

Heissel, J. A. *et al.* (2021). Testing, stress, and performance: how students respond physiologically to high-stakes testing. *Education Finance and Policy*, **16**, 183–208.

Högberg, B. *et al.* (2019). Consequences of school grading systems on adolescent health: evidence from a Swedish school reform. *Journal of Education Policy*, **36**, 84–106.

Högberg, B., Strandh, M. and 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.

Kerckhoff, A. (1995). Institutional arrangements and stratification processes in industrial societies. *Annual Review of Sociology*, **21**, 323–347.

Landstedt, E. and Gådin, K. G. (2012). Seventeen and stressed – do gender and class matter? *Health Sociology Review*, **21**, 82–98.

Legewie, J. and DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, **77**, 463–485.

Lee, M. and Larson, R. (2000). The Korean 'Examination Hell': long hours of studying, distress, and depression. *Journal of Youth and Adolescence*, **29**, 249–271.

Löfstedt, P. *et al.* (2020). School satisfaction and school pressure in the WHO European region and North America: an analysis of time trends (2002–2018) and patterns of co-occurrence in 32 countries. *Journal of Adolescent Health*, **66**, S59–S69.

Meyer, J. (1977). The effects of education as an institution. *American Journal of Sociology*, **83**, 55–77.

Mickelson, R. A. (1989). Why does Jane read and write so well? The anomaly of women's achievement. *Sociology of Education*, **62**, 47–63.

Montt, G. and Borgonovi, F. (2018). Combining achievement and well-being in the assessment of education systems. *Social Indicators Research*, **138**, 271–296.

OECD (2020). Gross domestic product (GDP) (indicator), available from: https://data.oecd.org/gdp/gross-domestic-product-gdp.htm [accessed 21 October 2020].

Pearlin, L. I. *et al.* (1981). The stress process. *Journal of Health and Social Behavior*, **22**, 337–356.

Pekkarinen, T. (2012). *Gender Differences in Education*. IZA Discussion Paper No. 6390. Institute of Labor Economics.

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: a meta-analysis. *Evaluation Review*, **43**, 111–151.

Reay, D. and Wiliam, D. (1999). 'I'll be a nothing': structure, agency and the construction of identity through assessment. *British Educational Research Journal*, **25**, 343–354.

Roberts, C. *et al.*; HBSC Study Group International. (2009). The Health Behaviour in School-aged Children (HBSC) study: methodological developments and current tensions. *International Journal of Public Health*, **54 Suppl 2**, 140–150.

Ross, C. E. and Mirowsky, J. (2006). Sex differences in the effect of education on depression: resource multiplication or resource substitution? *Social Science & Medicine*, **63**, 1400–1413.

Scheeren, L., van de Werfhorst, H. G. and Bol, T. (2018). The gender revolution in context: how later tracking in education benefits girls. *Social Forces*, **97**, 193–220.

Segool, N. K. *et al.* (2013). Heightened test anxiety among young children: elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools*, **50**, 489–499.

Sonmark, K. *et al.* (2016). Individual and contextual expressions of school demands and their relation to psychosomatic. *Child Indicators Research*, **9**, 93–109.

Sørensen, A. B. (1970). Organizational differentiation of students and educational opportunity. *Sociology of Education*, **43**, 355–376.

Tuominen-Soini, H. and Salmela-Aro, K. (2014). Schoolwork engagement and burnout among Finnish high school students and young adults: profiles, progressions, and educational outcomes. *Developmental Psychology*, **50**, 649–662.

van Hek, M., Buchmann, C. and Kraaykamp, G. (2019). Educational systems and gender differences in reading: a comparative multilevel analysis. *European Sociological Review*, **35**, 169–186.

Verger, A., Parcerisa, L. and Fontdevila, C. (2019). The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review*, **71**, 5–30.

Whitney, C. R. and Candelaria, C. A. (2017). The effects of no child left behind on children's socioemotional outcomes. *AERA Open*, **3**, 233285841772632.

Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data*. 2nd edn. Cambridge, MA: MIT.

Björn Högberg is a researcher and lecturer at Department of Social Work, and Centre for Demographic and Ageing Research (CEDAR), Umeå University, Sweden. His current research is focused on education policy, gender inequality, and health. He has also studied the interrelations between unemployment and insecure employment, and labour market institutions and education systems, among youth.

Daniel Horn is the director of the Institute of Economics and head of the Education and Labour Economics research unit at the Centre for Economic and Regional Studies (KRTK). Dr Horn is also affiliated with the Economics Institute of the Corvinus University of Budapest. He has conducted research on education inequality, tracking, vocational education, educational performance measurement issues, school-to-work transition, and gender-related topics.