INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

KRTK KTI WORKING PAPERS | KRTK KTI MŰHELYTANULMÁNYOK

High-stakes national testing, gender and school stress in Europe – A difference-in-difference analysis.

Björn Högberg – Dániel Horn

KRTK KTI WP – 2021/11

February 2021

https://www.mtakti.hu/wp-content/uploads/2021/02/CERSIEWP202111.pdf

INSTITUTE OF ECONOMICS, CENTRE FOR ECONOMIC AND REGIONAL STUDIES,
BUDAPEST, 2021

ABSTRACT

Outcomes related to the wellbeing of students are increasingly being recognized as valuable objectives for education systems. In this study, we ask if high-stakes testing affects school-related stress among students and if there are gender differences in these effects.

We combine macro-level data on high-stakes testing with survey data on more than 300,000 students aged 10-16 years in 31 European countries, from three waves (2002, 2006 and 2010) of the Health Behaviour in School-aged Children (HBSC) study. With variation in high-stakes testing across countries, years and grade levels, we use a quasi-experimental difference-in-differences (DD) design for identification of causal effects. We find that high-stakes testing increases self-reported school-related stress by almost 10 % of a standard deviation. This is primarily driven by a strong effect for girls, meaning that high-stakes testing increases the gender gap in school-related stress. The results are robust to a range of sensitivity analyses.

JEL codes: I14, I21, I24,

Keywords: high-stakes testing, stress, gender, quasi experiment

Björn Högberg
Department of Social Work, Umeå University, Sweden
Centre for Demographic and Ageing Research (CEDAR), Umeå University, Sweden
e-mail: bjorn.hogberg@umu.se.

Dániel Horn
Centre for Economic and Regional Studies, Institute of Economics, Hungary
Corvinus University of Budapest, Hungary
e-mail: daniel.horn@krtk.hu

Téttel bíró vizsgák, nemek és iskolai stressz Európában – egy különbség-a-különbségekben elemzés

Björn Högberg – Dániel Horn

ÖSSZEFOGLALÓ

A hallgatói jólét egyre inkább az oktatási rendszerek fontos célkitűzésévé válik. Ebben a tanulmányban azt nézzük meg, hogy a nagy téttel bíró tesztek befolyásolják-e az iskolai stresszt a tanulók körében, és vannak-e nemi különbségek ezekben a hatásokban.

Összevetjük a nagy téttel bíró tesztek makroszintű adatait 31 európai országban több mint háromszázezer 10-16 éves diák egészségmagatartásáról szóló felmérésének adataival 2002, 2006 és 2010-ből. A nagy téttel bíró tesztek országonkénti, évenkénti és évfolyamonkénti eltéréseinek felhasználásával kvázi-kísérleti különbség-a-különbségekben (DD) módszert alkalmazunk az ok-okozati hatások azonosítására. Eredményeink azt mutatják, hogy a nagy téttel bíró tesztek a szórás csaknem 10% -ával növelik az iskolai stresszt. Ezt elsősorban a lányokra gyakorolt erős hatás vezérli, ami azt jelenti, hogy a nagy téttel bíró tesztek növelik a nemek közötti szakadékot az iskolai stresszben. Az eredmények számos robosztussági teszt alapján is megbízhatóak.

JEL: I14, I21, I24,

Keywords: nagy téttel bíró vizsgák, stress, nem, kvázi kísérlet

High-stakes national testing, gender and school stress in Europe – A difference-in-difference analysis.

Björn Högberg[a*]

Dániel Horn[b]

[a] Department of Social Work, Umeå University, and Centre for Demographic and Ageing Research (CEDAR), Umeå University, Sweden

[b] Centre for Economic and Regional Studies, Institute of Economics, and Corvinus University of Budapest, Hungary

* Corresponding Author: Björn Högberg. Address: Department of Social Work Umeå University, SE-901 87, Umeå, Sweden. Email: bjorn.hogberg@umu.se. Telephone: +46 705937195

Acknowledgements

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

High-stakes national testing, gender and school stress in Europe – A difference-in-difference analysis.

Abstract

Outcomes related to the wellbeing of students are increasingly being recognized as valuable objectives for education systems. In this study, we ask if high-stakes testing affects school-related stress among students and if there are gender differences in these effects.

We combine macro-level data on high-stakes testing with survey data on more than 300,000 students aged 10-16 years in 31 European countries, from three waves (2002, 2006 and 2010) of the Health Behaviour in School-aged Children (HBSC) study. With variation in high-stakes testing across countries, years and grade levels, we use a quasi-experimental difference-in-differences (DD) design for identification of causal effects.

We find that high-stakes testing increases self-reported school-related stress by almost 10 % of a standard deviation. This is primarily driven by a strong effect for girls, meaning that high-stakes testing increases the gender gap in school-related stress. The results are robust to a range of sensitivity analyses.

JEL codes: I14, I21, I24,
Keywords: high-stakes testing, stress, gender, quasi experiment

Téttel bíró vizsgák, nemek és iskolai stressz Európában – egy különbség-a-különbségekben elemzés

Absztrakt

A hallgatói jólét egyre inkább az oktatási rendszerek fontos célkitűzésévé válik. Ebben a tanulmányban azt nézzük meg, hogy a nagy téttel bíró tesztek befolyásolják-e az iskolai stresszt a tanulók körében, és vannak-e nemi különbségek ezekben a hatásokban.

Összevetjük a nagy téttel bíró tesztek makroszintű adatait 31 európai országban több mint háromszázezer 10-16 éves diák egészségmagatartásáról szóló felmérésének adataival 2002, 2006 és 2010-ból. A nagy téttel bíró tesztek országonkénti, évenkénti és évfolyamonkénti eltéréseinek felhasználásával kvázi-kísérleti különbség-a-különbségekben (DD) módszert alkalmazunk az ok-okozati hatások azonosítására.

Eredményeink azt mutatják, hogy a nagy téttel bíró tesztek a szórás csaknem 10% -ával növelik az iskolai stresszt. Ezt elsősorban a lányokra gyakorolt erős hatás vezérli, ami azt jelenti, hogy a nagy téttel bíró tesztek növelik a nemek közötti szakadékot az iskolai stresszben. Az eredmények számos robosztussági teszt alapján is megbízhatóak.

JEL kódok: I14, I21, I24,
Kulcsszavak: nagy téttel bíró vizsgák, stress, nem, kvázi kísérlet

Introduction

Increasing academic outcomes have traditionally been the core aim of education policy. However, outcomes related to the wellbeing of students are increasingly being recognized as valuable objectives in themselves (Montt and Borgonovi, 2018; OECD, 2017). One such non-academic outcome is school-related stress.

School stress is associated with poor mental health and wellbeing among adolescent students, including higher levels of psychosomatic health complaints (Sonmark et al., 2016), depressive symptoms (Aanesen, Meland, and Torp, 2017; Barker et al., 2018), poor psychological wellbeing and sleep problems (Pascoe, Hetrick, and Parker, 2020), and suicidal ideation (Lee et al., 2006; Wang, 2016). These associations hold for different national and educational contexts (Cosma et al., 2020), and in cross-sectional as well as longitudinal studies (Aanesen, Meland, and Torp, 2017; Barker et al., 2018). Girls are more sensitive to school stress, and school stress is an important factor behind the gender gap in mental ill-health that opens up during adolescence (Aanesen, Meland, and Torp, 2017; Högberg, Strandh, and Hagquist, 2020). Since the onset of mental disorders often occurs during adolescence (Kessler et al., 2007), excessive school stress may have long-term harmful consequences, and conversely, preventive efforts that reduce stress may have large positive gains.

While a large body of literature has documented the negative consequences of school stress, relatively less attention has been paid to what causes school stress in the first place. The existing literature on school stress and related aspects such as test anxiety has mainly focused on the individual student and the immediate school context (Banks and Smyth, 2015; Sonmark et al., 2016; von der Embse et al., 2018), rather than on education policies and other institutional factors. One potentially important institutional factor in this regard is the testing policy of education systems. When students are asked about what they perceive as stressful in life, tests, examinations and marks are consistently ranked as among the most important stressors (Byrne, Davenport, and Mazanov, 2007). Also, reported levels of stress and/or anxiety increase close to major tests, especially for girls and when the stakes are high (Heissel et al., 2019; West and Sweeting, 2003). In line with this, girls are more reluctant to engage in competitive and high-stakes testing (Nekby, Skogman Thoursie and Vahtrik, 2015; Niederle and Vesterlund, 2011), and report higher stress in relation to these tests (Banks and Smyth, 2015; Kouzma and Kennedy, 2004; Landstedt and Gådin, 2012; Wiklund et al., 2012).

Despite a global rise in national high-stakes testing (Verger, Parcerisa, and Fontdevila, 2019), and a concomitant increase in school stress especially among girls (Cosma et al., 2020; Löfstedt et al., 2020), systematic studies of the effects of high-stakes testing policies on stress and related outcomes are still scarce (for recent exceptions, see Högberg et al. (2019) and Whitney and Candelaria (2017)). Using repeated cross-sectional survey data on more than 300 000 students in 31 European countries from 2002 to 2010, the aim of the present study is to investigate the effects of national high-stakes testing on school stress among adolescents, with a specific focus on gender differences.

We build upon and extend the existing literature on school stress in three ways. First, we combine the literature on gender differences in competitiveness and high-stakes testing (Niederle and Vesterlund, 2011) with the literature on gender differences in school stress. Second, we focus on the hitherto relatively neglected role of institutional factors at the level of education systems in generating stress. Third, using a quasi-experimental differences-in-differences design with multiple groups and time periods, we improve on causal inference, and are able to draw policy relevant conclusions on the impact of national testing policies on school stress.

Background and previous research

School stress

Stress is a potentially ambiguous concept, with different partly overlapping meanings (Putwain, 2007). We use the term *stressor* to designate external conditions or stimuli in the form of threats, challenges or demands, and *stress* to designate the subjective (psychological or physiological) *response* to these stimuli (Wheaton, 2013; Putwain, 2007). The conceptualization of *stress* used in this study is ultimately limited by the available data (see below), but in a general sense stress is understood as responses to demands that are perceived as unmanageable and threatening (cf. Lazarus and Folkman, 1984). 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.

Previous research

While research on individual level correlates of school stress or anxiety is extensive (see the meta analysis by von der Embse et al. (2018)), less is known about the role of institutional factors such as high-stakes testing. One strand of research has examined levels or perceptions of stress close to high-stakes tests, with most finding suggesting that stress is higher around these tests (Banks and Smyth, 2015; Heissel et al., 2019; Segool et al., 2013; West and Sweeting, 2003). Another strand of research has investigated the effects of changes in testing policies. Whitney and Candelaria (2017), in a study of the introduction of high-stakes testing tied to school accountability laws (No Child Left Behind) across US states, found some evidence of positive effects on self-reported anxiety, but no evidence of gender differences in these effects. Högberg et al. (2019), on the other hand, found rather substantial positive effects on self-reported stress of a reform that introduced teacher-assigned marks and large-scale national tests in lower grade levels in Sweden. These effects were slightly, but not significantly, stronger for girls. Also of relevance is Markowitz (2018), who found that introduction of high-stakes testing across US states increased school engagement in the short-run but decreased it in the long run. A crucial difference between these studies is that the tests studied by Whitney and Candelaria (2017) and Markowitz (2018) were primarily high-stakes for schools, while the marks and tests studied by Högberg et al., (2019) are high-stakes for the individual student.

Theoretical framework

The importance of schools and education systems for the wellbeing of students, including school stress, can be understood from the perspective of how schools sort students into two interlocking hierarchies (Elstad, 2010). First, schools sort students into a hierarchy of performance through various forms of assessment and evaluation. Second, since education and educational performance is the main mechanism through which individuals are allocated to social positions in the labor market, positions in the hierarchy of school performance is translated into positions in a hierarchy of social prospects. These two hierarchies combine to generate school stressors for students.

With regard to the first hierarchy, qualitative studies show that students, especially girls, view educational performance as an important part of their identity. A poor performance might therefore become a threat to this identity. When educational performance is linked to processes of social comparison, poor performance may also become a threat to social status, self-worth and self- esteem (Landstedt, Asplund, and Gillander Gådin, 2009; Låftman, Almquist, and Östberg, 2013; Putwain, 2007). Since tests and other measurements of performance makes students' relative performance explicit, they likely increase the salience of these forms of social comparison (Eccles, 1989).

Moreover, since studying is the principal means through which students can affect their measured performance, the performance hierarchy increases stressors related to the workload of students. Accordingly, a high workload, with many tests, time-consuming homework and tight deadlines, are considered to be among the most stressful aspects of school (Banks and Smyth, 2015; Kouzma and Kennedy, 2004; Lee and Larson, 2000; Smyth and Banks, 2012; Wiklund et al., 2012). Due to their perceived importance, high-stakes tests are likely to amplify these effects, more so than other forms of low-stakes assessments (Banks and Smyth, 2015; Putwain, 2009; Segool et al., 2013; Wang, 2016).

As regards the second hierarchy (social prospects), the important aspect in the context of this study is that adolescents and also children are highly aware of the importance of educational performance for their future opportunities in the education system and the labor market (Huan et al., 2008; Låftman, Almquist, and Östberg 2013). All education systems are sequentially structured and path-dependent at some level, such that progression to higher levels of the education system, or to specific programs or tracks (e.g. vocational or academic tracks) within these levels, is dependent on performance at previous levels (Breen and Jonsson, 2000). The sequential structure implies that school performance also at lower educational levels can have far-reaching consequences for future educational and labor market opportunities. High-stakes testing plays a crucial role in this sequential structure, as they are often used to award certificates and determine eligibility for progression to higher levels, and for sorting students into different study programs or tracks within these. Accordingly, several qualitative studies have found that students view their performance on high-stakes tests as crucial for their future life chances, and that this is one of the main reasons for why these tests are experienced as more stressful than other types of tests (Banks and Smyth, 2015; Denscombe, 2000; Landstedt and Gådin, 2012; Låftman, Almquist, and Östberg, 2013; Putwain, 2009).

Gender differences in stress

A consistent finding in research on school stress and test anxiety is that girls report substantially higher levels of stress, and that gender differences grows through adolescence (Aanesen, Meland, and Torp, 2017; Huan et al., 2008; Högberg, Strandh, and Hagquist, 2020; Sonmark et al., 2016; Östberg et al., 2015). This can be understood from the perspective of the previously discussed two school-related hierarchies. First, the identity, self-worth and self-esteem of adolescent girls tends to be more closely tied to school and school performance (Landstedt and Gådin, 2012; Schraml et al., 2011). Second, girls tend to have higher educational aspirations (van Houtte, 2017), and are more likely to enter in academic tracks or enroll in higher education for which they typically need good grades or test results. Also, women tend to have higher returns from education, such that the labor market opportunities for women are more dependent on their education level (Pekkarinen, 2012). Thus, education and educational performance is more important for the social prospects of girls and women than for men, meaning that high-stakes test may be “higher-stakes” tests for girls.

Another line of research shows that girls are less competitive, more risk averse, and more likely to underestimate their ability in educational tasks. High-stakes testing is among the most competitive situations that students engage in at school; accordingly, girls are more reluctant to engage in competitive or high-stakes testing and tend to underperform in high-stakes tests relative to their ability and to their performance in other forms of tests (Niederle and Vesterlund, 2011). In addition, Niederle and Yestrumskas (2008) find that women are more likely to choose challenging educational tasks when choices are flexible compared to when choices are irreversible and have appreciable consequences, as is the case with high-stakes tests. The results of Nekby, Skogman Thoursie and Vahtrik (2015) also suggest that female students prefer flexible choice architectures in testing situations. Hence, women are more likely to perceive educational choices or tests as threatening when the consequences of their performance are large and long-lasting.

Data and methods

National testing is by definition shared by all students in an education system and its effects may therefore be difficult to measure in a single country at a single point in time. With harmonized cross-country data, spanning over years and school grade levels, we can use variation in testing policies across countries, grade levels 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 survey of students in primary and lower secondary school in high- and middle-income countries, conducted in collaboration with the World Health Organization every four years since the 1980s. HBSC is among the most comprehensive international adolescent health surveys in terms of included countries and years, and has been extensively used in comparative research on school stress (Löfstedt et al., 2020; Sonmark et al., 2016). HBSC data is collected through a cluster sampling design, with schools or school classes as primary sampling units. To ensure cross-country comparability of the data, all participating countries must abide to the standardized research protocol for data collection and analysis. HBSC collects data on three age groups aged 11.5, 13.5 and 15.5 years on average, corresponding to three country-specific grade levels. Each country draws a sample of school classes in each group, with the minimum sample size per group set at 1500 students (see Roberts et al. (2009) for more information on the survey methodology). Response rates vary across countries and surveys, with average response rates around 70 % (HBSC, 2021). We use the maximum amount of countries and survey years (2002, 2006, 2010) for which we have matching country-level data on of national testing policies (see below), resulting in a dataset with more than 300,000 students, nested in 31 European countries and three survey years.

Country-level data

We use data from Eurydice to identify national testing policies of countries (Eurydice, 2009). Eurydice is a European Union network with the task to provide comparable information on European education systems. Eurydice data on national testing has been validated in previous comparative studies on European education systems (Braga, Checchi, and Meschi, 2014). Since

Eurydice data on high-stakes testing are not available after 2009/2010 we cannot use the latest HBSC surveys of 2014 and 2018. Eurydice only covers national tests in ISCED levels 1 (primary education) and 2 (lower secondary education). Eurydice also only cover national tests (regional in the case of Belgium and Great Britain), defined as tests standardized by national education authorities or top-level authorities for education. Tests at lower administrative levels and non-compulsory tests are not included.

Eurydice data has two critical advantages given the aims of this study. Firstly, Eurydice distinguishes between three types of national tests depending on their purpose: (1) Tests used for “taking decisions about the school career of pupils”, (2) tests used for “monitoring schools and/or the education system” and (3) test used for “identifying individual learning needs”. The first category includes tests that are used to determine eligibility to higher levels of the education system, to determine track placement, or to determine grade retention. The second category includes test used to evaluate schools or monitor the education system. The third category includes tests used to identify the degree to which students reach stated learning goals, but have no direct consequences on the continuation of students within the education system.

The analytical focus of the study is on the effects of high-stakes testing. Tests can be high-stakes either for students or the school (when results of tests are tied to sanctions and rewards for the school, as in outcome-based accountability systems). Of the three categories distinguished by Eurydice, only the first refer to tests that are high stakes for the student, since the results have direct consequences for their future educational trajectories. The second category includes tests that may be high-stakes for schools, but also tests that are performed solely for administrative purposes. Tests included in the third category are primarily for formative purposes and do not have high-stakes for any actor. We will use the first category as our main variable of interest.

The second advantage with Eurydice data is that they provide information on the year in which the national test was first implemented and the grade level in which students take the test. With repeated cross-sectional data at the individual level from different time points and grade levels, we can use variation across time and grade levels in addition to variation across countries for estimation. It should be noted that we only have individual data on three grade levels, meaning that we cannot identify effects of tests held in other grade levels.

Dependent variable

We conceptualize school stress as perceptions of excessive demands related to school. HBSC data contain one item on school stress that is available in all countries and years. The item intends to measure the global feeling of being pressured by schoolwork, which includes work at school and homework. The question asked is *"How pressured do you feel by the schoolwork you have to do?"*, with answers ranging from "Not at all" (0) to "A lot" (3). While a more detailed and comprehensive measure, directly focusing on the role of testing and not school work in general, would have been desirable, the item has some advantages for the purpose of this study. Along 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 adolescent health (Sonmark et al., 2016), and has therefore been used *in lieu* of the full subscale to track levels and correlates of schools stress across countries (Löfstedt et al., 2020). The item has been extensively used in previous studies on school stress, and is associated with both school satisfaction and health across European countries (Aanesen, Meland, and Torp, 2017; Högberg, Strandh, and Hagquist, 2020). For these reasons, we believe that the item may serve as a valid proxy for school stress. In our main model we use the variable as continuous as there is large variance in between categories in the overall sample as well as within countries (see figures S1 and S2 in the appendix).

Covariates at the individual level

An important individual level variable is the grade level of the student. HBSC collects data on three age groups aged on average 11.5, 13.5 and 15.5 years (henceforth 11, 13 and 15 years). This corresponds to three grade levels, which are the typical grade levels for students in this age group in the given countries. In some countries, students in the relevant age group may be spread over more than one grade level due to high rates of grade retention, in which case part of the sample may be drawn from other than the typical grade level (Roberts et al., 2009). The average retention rate in lower secondary education in Europe is around 10 %, with substantial variation across countries (Eurydice, 2011). However, most participating countries in HBSC only draw samples from the age-typical grade levels. Thus, the share of the sample for which the sampled age group is spread over more than one grade level is substantially lower than 10 %, and only about 1 % of the sample differ by more than one 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 since the share of students repeating a grade in secondary school is very similar in countries with or without high stakes testing (6.4 % vs 7.7 %). Moreover, even if this difference is considered important, because low-ability students generally show higher test anxiety (von Embse, 2018), lower number of grade repeaters in high-stakes testing countries would also probably indicate lower levels of stress; hence our estimates would be downwardly biased.

We include a rich set of individual-level covariates. Firstly, to account for compositional differences of the student population across countries and years, we control for age, gender, whether the student lives with both parents, and the student's perception of the economic standing of the family (a proxy for social background). We label these as “demographic” controls. We also control for a range of individual characteristics that may be associated with school stress. These include the quality of child-parent relationships, the student's opinion of his/her body, frequency of binge drinking, frequency of physical activity, one indicator measuring how often the respondent has been bullied in school, one indicator measuring how often the respondent has been in a physical fight, and one indicator measuring how often the respondent has been injured in some way (Banks and Smyth, 2015; Högberg, Strandh, and Hagquist, 2020; Låftman & Modin, 2012). We label these as “additional student level” controls. Unlike the student background characteristics, the latter set of covariates may potentially be caused by stress related to high-stakes testing. In other words, from a theoretical perspective it is not obvious whether they should be regarded as common cause or unwanted mechanism confounders of the relationship between testing and stress.

Covariates at the country level

We include three time varying country-level covariates: gross domestic product (GDP) per capita, country-level economic inequality as measured by the GINI index, and the youth unemployment rate (data from Eurostat (2020a; 2020b) and OECD (2020)). While studies on macro-level determinants of school stress are, to the best of our knowledge, lacking, research shows that these indicators are related to student's wellbeing more broadly (Elgar et al., 2015; Johansson et al., 2019).

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

Analytical strategy

We use a quasi-experimental difference-in-difference (DD) design, with multiple groups and periods. In the most basic setting, a DD-design compares two groups, one treatment group and one control group, before and after some event ("treatment") that only affects the treatment group. The effect of the treatment on a certain outcome is then the difference in the change over time in the outcome in the treatment group and the equivalent change in the control group. This design controls for all time-invariant differences between the two groups, as well as for all time trends that are common to both groups. We extend this basic setting using multiple groups (countries) and periods (survey years), and also, since we have data on the grade level of students, multiple comparison groups (grade levels) within countries, resulting in a difference-in-difference-in-difference (DDD) design (Imbens and Wooldridge, 2009). With multiple groups and periods, DDD-designs are best analyzed within a regression framework. Specifically, we estimate a three-way fixed effects regression model of the following form:

$$\gamma_{igcy} = \beta_0 + \beta_1 C_c + \beta_2 G_g + \beta_3 Y_y + \beta_4 T_{gcy} + \beta_5 X_{igcy} + \beta_6 Z_{cy} + \varepsilon_{icyg} \qquad \text{(Eq. I)}$$

where *i* stands for the individual student, *g* for grade level, c for country and *y* for survey year. A full set of dummy variables is included for each of the groups and time periods. Dummy variables for countries are denoted by C_c, and captures all time invariant differences across countries. Dummy variables for grade levels are denoted by G_g, and captures all time invariant differences across grade levels. Dummy variables for survey years are denoted by Y_y, and captures time trends that are invariant across countries and grades. We also estimate fully saturated models, with a full set of two-way interactions between the dummies for country, survey year, and grade level, that is $C_c * G_g$, $C_c * Y_y$ and $G_g * Y_y$. X_{igcy} is a vector of individual-level controls that includes demographic controls as well as other individual characteristics, as described previously. Z_{cy} is a vector of time-varying country-level covariates, as described previously (these drop out from the fully saturated models). $\varepsilon_{icyg}$is an individual specific error term.

The main explanatory variable – national high-stakes tests with consequences for the school career of students – is denoted by T_{gcy}. This dummy variable is coded 1 for students in countries, grade levels and survey years in which such a national test is held, and 0 otherwise. Thus, β_4 gives the effect of national high-stakes testing and is the main parameter of interest. The estimation of β_4 is based on variation within countries across grades and years. To study gender differences, we introduce an interaction term between the high-stakes testing indicator and gender in the last, fully-saturated model.

We use a linear least square dummy variable (LSDV) 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 the treatment assignment, which is grade levels within countries. Because the treatment does not vary between schools within countries, we do not cluster standard errors at the school level.

We use wild cluster bootstrap for inference, since this is more conservative and tend to perform better in finite samples compared to default cluster robust standard errors (MacKinnon and Webb, 2020). Cluster robust standard errors, and even more so wild cluster bootstrap, perform well when the number of clusters is not too small (less than around 40), and when the number of observations per cluster is of approximately equal size (Cameron and Miller, 2015). With 93 clusters, and little variation in cluster size, the data at hand should be enough for reliable inference. Cluster robust standard errors are also heteroskedastic consistent, which is of importance as we use linear regression with an ordinal scale outcome variable (Cameron and Miller, 2015).

The main assumption required for drawing causal conclusions from a difference-in-difference analysis is the parallel trends assumption (Angrist and Pischke, 2009; Wing, Simon, and Bello-Gomez, 2018). This implies that in the absence of treatment, the differences in outcomes between treatment and control groups would be constant over time. Thus, the parallel trends assumption implies that the control group can serve as a counterfactual for what would happen to the treatment group had they not received treatment (in our case: high-stakes tests), or in other words, that confounding is constant across treatment and control groups (Wing, Simon, and Bello-Gomez, 2018). Since the parallel trends assumption refers to counterfactual outcomes that cannot be observed, it cannot be tested directly. The credibility of the assumption can, however, be probed in different ways, an issue we return to in the results section. A related assumption is that the composition of the treatment and control groups does not change over time (Angrist and Pischke, 2009). We address this by including a rich set of student background characteristics, and checking for covariate balance in supplementary analyses. 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 (SUTVA) of the counterfactual causal model, that is, that treatment of one unit does not affect outcomes of non-treated units. For reasons laid out later in the results-section, we do not believe that such spillover effects are a major problem.

Results

Table 1 provides information on the presence of national high-stakes testing in the various countries, grades and, when relevant, the year of introduction or abolishment. Since the terms used to denote grade levels differ across countries depending on school starting age, we express the grade level of the test using the typical age for students at that grade level. We only list national and compulsory high-stakes tests, as defined by Eurydice (Eurydice, 2009). Less than half of the countries have no high-stakes tests and a few have tests in grade levels not covered in HBSC. In most countries, high-stakes tests are taken at age 15-16, which typically corresponds to the last year before upper secondary school. Seven countries (Bulgaria, Belgium (Walloon region), Germany, Iceland, Italy, Norway and Romania) have introduced or abolished tests between 2002 and 2010.

Table 1. Information on national high-stakes testing across countries, years and grade levels.

Country	High-stakes testing for 11 year olds	High-stakes testing for 13 year olds	High-stakes testing for 15 year olds	High-stakes testing, other grade levels	National tests for other purposes
Austria					
Belgium, Flemish Region					
Belgium, Walloon region	X (from 2009)				X
Bulgaria				X	X (from 2006)
Czech Republic					
Denmark			X		X (from 2010)
Estonia			X		
Finland					
France					X
Germany			X (from 2009)		
Greece					
Hungary					X
Iceland			X (until 2009)		X
Ireland			X		X
Italy		X (from 2008)			X
Latvia			X		X
Lithuania					
Luxembourg	X				X
Malta			X		
Netherlands	X				
Norway			X (from 2004)		X (from 2004)
Poland			X		X
Portugal				X	X
Romania		X (from 2007)			X
Slovakia					X
Slovenia					X
Spain					
Sweden			X		X
England					X
Scotland			X		X
Wales					

Only compulsory and national tests used for “taking decisions about the school career of pupils” included in the high-stakes category. 11 years include students aged 11-12; 13 years include students aged 13-14 years; 15 years include students aged 15-16 years. Tests for other purposes include compulsory and national tests in any grade level, used for “Identifying individual learning needs” and “Monitoring schools and/or the education system”.

We present the main results in Table 2. The table includes six models, with stepwise inclusion of country and student level covariates. Our baseline model 1 includes fixed effects for country, year and grade level in addition to the treatment variable. High-stakes testing is associated with approximately 0.083 scale steps higher levels of stress. Since the mean and standard deviation of school stress in the total sample is 1.26 (mean) and 0.89 (standard deviation), this roughly corresponds to an almost 10 % standard deviation increase in stress. Model 2 includes the same covariates, but restricts the sample to complete cases, so as to enable comparison with models with additional covariates. The coefficient for high-stakes testing is substantially unchanged.

Model 3 adds the country-level covariates as well as the individual-level demographic controls, while in model 4 the additional student-level covariates are added. The coefficient for national high-stakes testing is stable across these specifications, though reduced slightly in model 4 (from 0.077 to 0.069). Model 5 is a fully saturated model with a full set of two-way interactions between dummy variables for country, survey year and grade level. The coefficient for high-stakes testing becomes somewhat larger, 0.096 compared to 0.069.

All in all, we see that high stakes testing has a statistically significant and non-negligible effect on school stress. Students in countries, grades and years, where high stakes testing is present, report almost 10 % of a standard deviation higher school stress than their counterparts with no high-stakes testing. This effect is an average effect for both genders, but based on previous research we expect the effect for girls to be higher.

Model 6 investigates gender differences in the effect of high-stakes testing, by interacting high-stakes testing with gender in the fully saturated model. It is apparent that girls suffer more from high-stakes testing than boys. The effect for boys is only 0.046 and not significant, while girls report an additional 0.097 points higher school stress, resulting in a total effect of 0.143 (0.046 + 0.097). The interaction term is similar in size to the average effect of high-stakes testing in model 5, which means that the gender gap in the effect of testing is as large as the stress-gap between high-stakes testing and non-high stakes testing countries. In short, the effects in models 1-5 are largely driven by the effects of testing on girls.

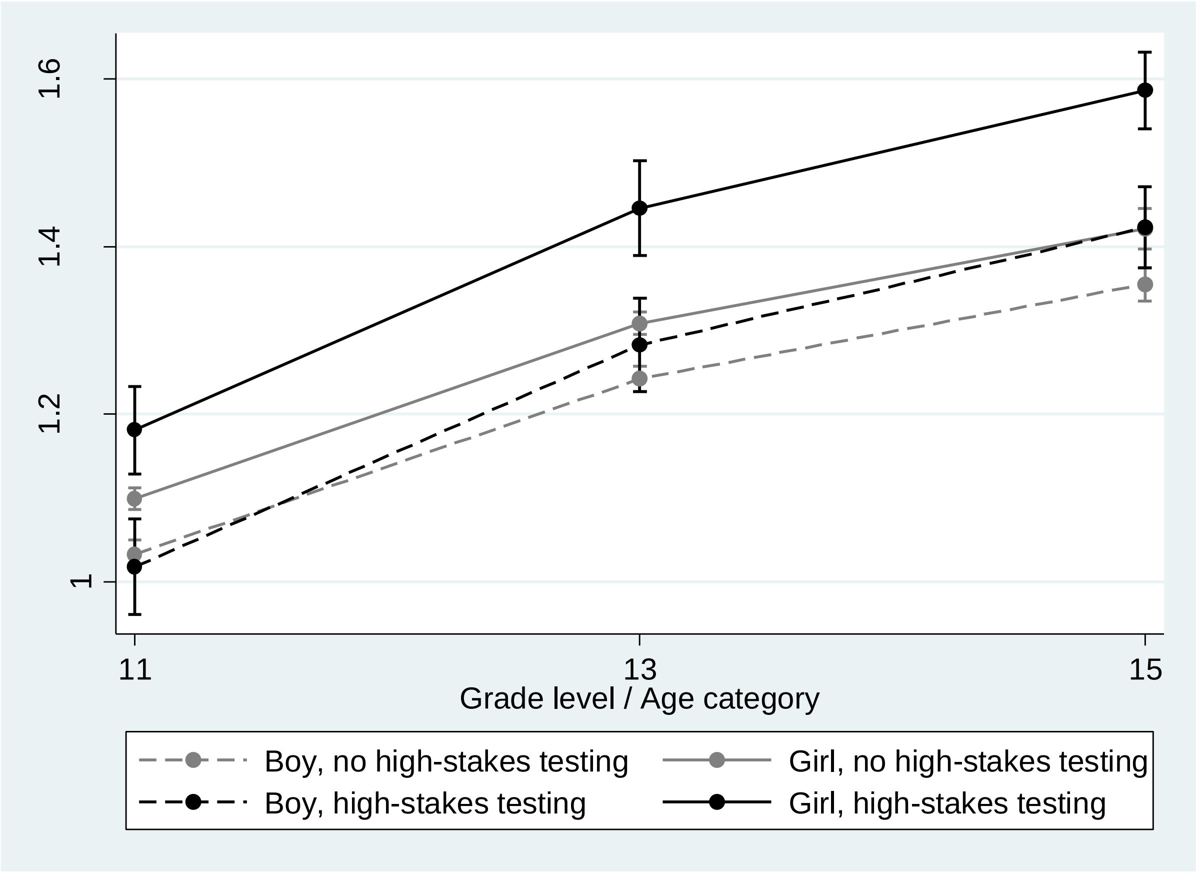
Table 2. Linear regression models with school stress as the outcome.

	m1	m2	m3	m4	m5	m6
	b	b	b	b	b	b
High-stakes testing	0.083* (0.035)	0.075* (0.034)	0.077* (0.035)	0.069* (0.033)	0.096*** (0.016)	0.046 (0.026)
Girl			0.061*** (0.014)	0.080*** (0.009)	0.080*** (0.013)	0.066*** (0.014)
High-stakes testing * Girl						0.097* (0.041)
Intercept	0.861	0.837	0.518	0.118	0.575	-0.585
Country, year, grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Two-way interactions: country, year, grade					Yes	Yes
Country covariates			Yes	Yes		
Demographic covariates			Yes	Yes	Yes	Yes
Additional student level covariates				Yes	Yes	Yes
sample	full	complete cases	complete cases	complete cases	complete cases	complete cases
N individual level	416,003	325,176	325,176	325,176	325,176	325,176
N country-grade level	93	93	93	93	93	93

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Besides between-gender we can also investigate between-grades heterogeneity of treatment as the literature suggests that stress grows during adolescence. Figure 1 below depicts the conditional treatment effects for the three grade levels and for the two genders, estimated from a regression model including two two-way interaction terms between the high-stakes testing indicator and, respectively, grade level and gender (exact coefficients in Table S3). We see that the average effects of model 5 is driven by gender as well as by grade differences. At age 11, there is no effect of high-stakes testing for boys (dashed black vs. dashed gray line), while girls in high-stakes testing countries have around 0.08 scale points higher levels of stress (solid black vs. solid gray line). At age 13, boys in high-stakes testing countries are somewhat (0.04 scale points) more stressed than boys in countries with no high-stakes testing, while the corresponding gap for girls grows to 0.14 scale points. At age 15, these gaps grow further, to around 0.07 (boys) and 0.16 (girls) scale points. Thus, effects are strongest in higher grade levels, possibly due to the forthcoming transitions to upper secondary school.

Figure 1 – Conditional effects of high-stakes testing, by grade level and age category



Spikes indicate 95 % confidence intervals.

Supplementary and sensitivity analyses

We test the credibility of the key assumptions of the research design in several ways. A standard way to test the parallel trends assumption is to examine the pre-treatment trends in the treatment and control groups. In this study, we only have three time-periods and most countries have introduced high-stakes test previous to our earliest available individual-level data in 2002. Of the seven countries that has introduced or abolished high-stakes testing during the study period, only one (Italy) has data on more than one pre-treatment period, while the others either introduced tests between 2002 and 2006 or did not participate in HBSC in 2002. Given these data restrictions, comparisons of pre-treatment trends are hardly informative. We do, however, address this issue in our fully saturated model by including the country and survey year interactions (i.e. introducing flexible country-specific time trends).

In a second set of sensitivity tests, we ran a set of placebo tests. In the first placebo test we re-estimated the preferred fully-saturated models (models 5 and 6) with lead values for the treatment indicator. Specifically, we manipulated the high-stakes testing indicator so that the grade level at which the test is conducted is moved two grade levels “upwards”, meaning that – for instance – the indicator is coded 1 for 13-year old students in countries with tests for 11-year old students, and so on. Results (column 1 and 2 in Table S4 in the online appendix) show that this manipulated variable has no effect on school stress.

A related placebo test is to estimate the fully-saturated models, but replace the high-stakes testing indicator with an equivalent indicator for national testing for other purposes that are not high-stakes for individual students (“monitoring schools and/or the education system” and “identifying individual learning needs”). We expect that these low-stakes tests should not affect stress. Results (column 3 and 4 in Table S4 in the online appendix) show that national testing for other purposes has no significant effect on stress.

We have also re-estimated the fully-saturated models while excluding countries with high rates of grade retention. This is because the correspondence between age group and grade level may be weaker when grade retention is more common. Belgium (Walloon region), France, Luxembourg, Portugal, Spain all have grade retention rates exceeding 20 % in lower secondary school (Eurydice, 2011). Excluding these countries does not affect the results in a substantial way (column 5 and 6 in Table S4 in the online appendix).

As a further robustness check, we have recoded the school stress variable into a dummy variable (0 = “Not at all” or “A little” stressed; 1 = “Some” or “Very” stressed) and run linear probability models on this variable. The results are substantially very similar (column 7 and 8 in Table S4 in the online appendix).

Since our identification strategy relies on variation over time (survey years) as well as across grade levels for estimation, we have also re-estimated the models using these two sources of variation separately, using either variation across survey years within grade levels, or variation across grade levels within survey years. With six separate survey years and grade levels, and two models (with and without interaction with gender), this results in 12 different models. Results are reassuring, in that the effects of high-stakes testing are, with one exception, consistently positive, though not consistently significant, in all 12 models (Tables S5 and S6 in the online appendix). The exception is the gender interaction effect, which is negative for the 13-year old sample when solely based on variation across survey years. Overall, these analyses suggest that variation across grade levels contributes relatively more to the total effect than does variation across survey years.

A further assumption of our analysis is that the compositions of the treatment and control groups do not change over time (or change similarly). This is partially addressed through the inclusion of covariates in the models. A more formal test is to check for covariate balance across treatment and control groups, which we do by re-estimating the fully-saturated models but replacing school stress with each of the individual-level covariates (one at a time) as the dependent variable. Among the 11 covariates, frequency of physical fighting was significantly associated with testing, and frequency of physical fighting and body image with the interaction between testing and gender (Tables S7 and S8 in the online appendix). However, with 11 variables in two different models (i.e. 22 focal coefficients), and a 5 % significance level, this may well be due to chance.

In addition to statistical tests, the credibility of the key assumptions may also be probed on theoretical grounds. A potential threat is reverse causality, in that high-stakes testing are introduced in countries, grades and years when reported school stress is high. Available data suggests that high-stakes testing in Europe has primarily been motivated by the need to allocate students, or to ensure that students meet learning goals (Eurydice, 2009; Verger, Parcerisa, and Fontdevila, 2019). Although some countries might have abolished or refrained from introducing tests due to concerns about stress, this would make our estimates a lower bound of the real effects. The reverse, that countries introduce tests in response to low stress, seems unlikely. On the whole, we do not believe that such reverse causality is major source of bias.

The data at hand is not sufficient to directly test the assumption of no spillover effects. Such spillover effects would be present if students in grades with no high-stakes testing were affected by the stress experienced by students in grades with tests. Qualitative research has shown that stress related to marks and tests may be “contagious”, and spread across students due to social comparisons (Låftman, Almquist, and Östberg, 2013). Spillover effects would also be present if the anticipation of tests in coming grade levels are perceived as stressful by students. The most likely scenario would then be that students in grades with no testing would experience higher stress due to upcoming tests, which would bias the estimate of the effect of testing downwards, making our estimates lower bounds. However, the previously discussed placebo test using lead values (Table S3) suggests that anticipation effects are not a major issue.

Discussion and conclusions

The aim of this study was to investigate the effects of national high-stakes testing on school stress among students, with a specific focus on gender differences. Results showed that high stakes testing increased self-reported school-related stress by almost 0.1 scale points, or almost 10 % of a standard deviation. This average effect was primarily driven by the effects for girls, with high-stakes testing substantially increasing the gender gap in stress. The effects were also stronger in higher grade levels (15 year olds), indicating that high-stakes testing in relation to the transition to upper secondary school may be especially stressful. The average effect size is comparable to the average gender gap in stress. The results were robust to a range of sensitivity analyses, including placebo tests using lead values or indicators of low-stakes tests, exclusion of countries with potentially lower quality data, and tests of covariate balance. We have also ruled out reverse causality and spillover effects.

Positive effects of high-stakes testing on stress are consistent with qualitative and quantitative studies showing a rise in stress close to high-stakes tests (Banks and Smyth, 2015; Heissel et al., 2019; West and Sweeting, 2003). Stronger effects for girls is also consistent with qualitative findings showing that girls tend to be more sensitive to the evaluation of school performance (Landstedt and Gådin, 2012), as well as with experimental research showing that girls are more reluctant to engage in competitive or high-stakes testing, partly because of lower competitiveness and greater risk-aversion (Nekby, Skogman Thoursie and Vahtrik, 2015; Niederle and Vesterlund, 2011). Stronger effects at higher ages are also consistent with previous research (Aanesen et al., 2017; Högberg et al., 2020; Sonmark et al., 2016).

However, both our main findings – the positive average effects (for both gender combined), as well as the stronger effect for girls in the interaction analysis – are only partly in line the results of Whitney and Candelaria (2017) and Markowitz (2018), both of whom studied effects of high-stakes testing in the US. The estimated 10 % of a standard deviation is larger than the comparable estimates on anxiety in Whitney and Candelaria (2017), which were not significant when adjusting for multiple hypothesis testing. And while 10 % of a standard deviation is close to the estimates on school engagement in Markowitz (2018), Markowitz only found that testing had a negative effect on school engagement in the long run. We believe that the main difference between our findings and theirs is that both Whitney and Candelaria and Markowitz investigated test that were primarily high-stakes for schools, not for students.

Limitations

The results of this study should be interpreted in light of its limitations. Eurydice only cover national tests. If national tests are correlated with other (e.g. regional) tests that are nonetheless experienced as high-stakes by students, this may lead to bias. Related to this, there may be a temporal mismatch between the date of data collection 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 more noisy and lead to attenuation bias, as students will be less stressed when tests are temporally distant. Moreover, we could only measure stress with a single indicator. While this indicator has shown desirable properties in previous studies (Sonmark et al., 2016), a comprehensive set of items, capturing different aspects of school stress, would have been desirable. In addition, stress was only measured through self-reports, and results might have been different if objective measures, such as cortisol levels or other biomarkers (Heissel et al., 2019; Östberg et al., 2015), were used.

Implications

Student wellbeing is a major goal for education policy (OECD, 2017). We have shown that national high-stakes testing increases self-reported school stress, and thus possibly as consequence mental health problems (Wang, 2016). Thus, the findings of this study have important implications for the optimal design of education systems (Montt and Borgonovi, 2018). Policymakers that value student wellbeing, and gender equality in wellbeing, would be advised to consider alternatives to high-stakes testing, or ways to lessen the stress caused by existing testing. In cases where high-stakes testing is used for secondary level placements, a less rigid sequential structure in the education system could make the tests less high stakes. This could involve making it easier to change between educational tracks or programs, or introducing second chance opportunities for students who fail at specific critical junctures. Such opportunities may alleviate stress, as students' performance at any given time will be less consequential for their future opportunities.

The findings of this study also have implications for gender equality in education systems. It is well established that higher levels of stress of anxiety is associated with worse performance on educational tests (von der Embse et al., 2018). If high-stakes testing increases stress, performance on such tests may confound ability with stress resilience or competitiveness, and thus provide a poor signal of underlying ability. As stress responses to high-stakes testing vary systematically across genders, high-stakes testing may generate unwarranted gender differences in observed school performance. From this perspective, the finding that girls become more stressed by high-stakes tests, and also tend to underperform in these (Niederle and Vesterlund, 2011), suggests that girls may be disadvantaged by education systems that sort students to different tracks, programs or selective schools based their test scores. This conclusion would be in line with Niederle (2017), who emphasize that educational institutions that reward competitiveness can disadvantage girls and women, and that more flexible choice or sorting mechanisms may be favorable in this regard.

Research ethics statement

The research was conducted using publicly available data and did not involve any human subjects. Ethical approval of the data collection was granted by ethic committees in the respective participating countries.

Acknowledgements

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

References

Aanesen, Fiona, Eivind Meland, and Steffen Torp. (2017). Gender differences in subjective health complaints in adolescence: The roles of self-esteem, stress from schoolwork and body dissatisfaction. *Scandinavian Journal of Public Health 45*(4) 389-396.

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge. (2017). When Should You Adjust Standard Errors for Clustering? *NBER Working Paper No. w24003*.

Angrist, Joshua and Jörn-Steffen Pischke. (2009). *Mostly harmless econometrics: an empiricist's companion*: Princeton : Princeton University Press.

Banks, Joanne, and Emer Smyth. (2015). 'Your whole life depends on it': academic stress and high-stakes testing in Ireland. *Journal of Youth Studies 18*(5) 598-616.

Barker, Erin T., Andrea L. Howard, Rosanne Villemaire-Krajden, and Nancy L. Galambos. (2018). The Rise and Fall of Depressive Symptoms and Academic Stress in Two Samples of University Students. *Journal of Youth and Adolescence 47*(6) 1252-1266.

Wheaton, Blair, Marisa Young , Shirin Montazer, and Katie Stuart-Lahman. (2013). Social Stress in the Twenty-First Century In C. S. Aneshensel, J. C. Phelan, and A. Bierman (Eds.), *Handbook of the Sociology of Mental Health* (pp. 299-323). Dordrecht: Springer Netherlands.

Braga, Michela, Daniele Checchi, and Elena Meschi. (2014). Educational policies in a long-run perspective. *Economic Policy 28*(73) 45-100.

Breen, R. and Jonsson J.O. (2000). Analyzing educational careers: a multinomial transition model. *American Sociological Review*. 65, 754–72.

Byrne, Don. G., S. C. Davenport, and Jason Mazanov. (2007). Profiles of adolescent stress: The development of the adolescent stress questionnaire (ASQ). *Journal of Adolescence 30*(3) 393-416.

Cameron, Colin A., and Douglas L. Miller. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources 50*(2) 317-372.

Cosma, Alina, Gonneke Stevens, Gina Martin, Elisa L. Duinhof, Sophie D. Walsh, Irene Garcia-Moya, . . . Margaretha de Looze. (2020). Cross-National Time Trends in Adolescent Mental Well-Being From 2002 to 2018 and the Explanatory Role of Schoolwork Pressure. *Journal of Adolescent Health 66*(6, Supplement) S50-S58.

Denscombe, Martyn. (2000). Social Conditions for Stress: Young people's experience of doing GCSEs. *British Educational Research Journal 26*(3) 359-374.

Eccles, Jacquelynne, and Carol Midgley (1989). Stage/environment fit: Developmentally appropriate classrooms for early adolescents. In R. E. A. C. Ames (Ed.), *Research on motivation in education* (pp. 139-186). San Diego, CA: : Academic Press.

Elgar, Frank J., Timo-Kolja Pförtner, Irene Moor, Bart De Clercq, Gonneke W. J. M. Stevens, and Candace Currie. (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*(9982) 2088-2095.

Elstad, Jon Ivar. (2010). Indirect health-related selection or social causation? Interpreting the educational differences in adolescent health behaviours. *Social Theory & Health 8*(2) 134-150.

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

Eurostat (2020b). Unemployment by sex and age [une_rt_a], Less than 25 years, percentage of active population. Retrieved October 21 2020 (http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do)

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

HBSC. (2021). Publications: International Reports. Retrieved October 14 2020

Heissel, Jennifer A., Emma K. Adam, Jennifer L. Doleac, David N. Figlio and Jonathan Meer. (2019). Testing, Stress, and Performance: How Students Respond Physiologically to High-Stakes Testing. *Education Finance and Policy*, (online publication)

Huan, Vivien S., Yeo Lay See, Rebecca P. Ang, and Chong Wan Har. (2008). The impact of adolescent concerns on their academic stress. *Educational Review 60*(2) 169-178.

Högberg, Björn, Joakim Lindgren, Klara Johansson, Mattias Strandh, and Solveig Petersen. (2019). Consequences of school grading systems on adolescent health: evidence from a Swedish school reform. *Journal of Education Policy* 1-23.

Högberg, Björn, Mattias Strandh, and Curt Hagquist. (2020). Gender and secular trends in adolescent mental health over 24 years – The role of school-related stress. *Social Science & Medicine* 112890.

Imbens, Guido W., and Jeffrey M. Wooldridge. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature 47*(1) 5-86.

Johansson, Klara, Solveig Petersen, Björn Högberg, Gonneke W. J. M. Stevens, Bart De Clercq, Diana Frasquilho, . . . Mattias Strandh. (2019). The interplay between national and parental unemployment in relation to adolescent life satisfaction in 27 countries: analyses of repeated cross-sectional school surveys. *BMC Public Health 19*(1) 1555.

Kessler, Ronald C., G. Paul Amminger, Sergio Aguilar-Gaxiola, Jordi Alonso, Sing Lee, and T. Bedirhan Ustün. (2007). Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry 20*(4) 359-364.

Kouzma, Nadya M., and Gerard A. Kennedy. (2004). Self-Reported Sources of Stress in Senior High School Students. *Psychological Reports 94*(1) 314-316.

Landstedt, Evelina, Kenneth Asplund, and Katja Gillander Gådin. (2009). Understanding adolescent mental health: the influence of social processes, doing gender and gendered power relations. *Sociology of Health & Illness 31*(7) 962-978.

Landstedt, Evelina, and Katja Gillander Gådin. (2012). Seventeen and stressed – Do gender and class matter? *Health Sociology Review 21*(1) 82-98.

Lee, Margaret T.Y., Betty P. Wong, Bonnie W.-Y. Chow, and Catherine McBride-Chang. (2006). Predictors of Suicide Ideation and Depression in Hong Kong Adolescents: Perceptions of Academic and Family Climates. *Suicide and Life-Threatening Behavior 36*(1) 82-96.

Lee, Meery, and Reed Larson. (2000). The Korean ‘Examination Hell’: Long Hours of Studying, Distress, and Depression. *Journal of Youth and Adolescence 29*(2) 249-271.

Låftman, Sara Brolin, Ylva B. Almquist, and Viveca Östberg. (2013). Students' accounts of school-performance stress: a qualitative analysis of a high-achieving setting in Stockholm, Sweden. *Journal of Youth Studies 16*(7) 932-949.

Låftman, Sara Brolin, and Bitte Modin. (2012). School-performance indicators and subjective health complaints: are there gender differences? *Sociology of Health & Illness 34*(4) 608-625.

Lazarus, R.S. and Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer.

Löfstedt, Petra, Irene García-Moya, Maria Corell, Carmen Paniagua, Oddrun Samdal, Raili Välimaa, . . . Mette Rasmussen. (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*(6, Supplement) S59-S69.

Markowitz, Anna J. (2018). Changes in School Engagement as a Function of No Child Left Behind: A Comparative Interrupted Time Series Analysis. *American Educational Research Journal 55*(4) 721-760.

Montt, Guillermo, and Francesca Borgonovi. (2018). Combining Achievement and Well-Being in the Assessment of Education Systems. *Social Indicators Research 138*(1) 271-296.

Nekby, Lena, Peter Skogman Thoursie, and Lars Vahtrik. (2015), Gender differences in examination behavior. *Economic Inquiry 53* 352-364.

Niederle, Muriel. (2017). A Gender Agenda: A Progress Report on Competitiveness. *American Economic Review 107*(5) 115-119.

Niederle, Muriel and Alexandra H. Yestrumskas (2008). Gender Differences in Seeking Challenges: The Role of Institutions. *NBER Working Paper No. 13922*.

Niederle, Muriel, and Lise Vesterlund. (2011). Gender and Competition. *Annual Review of Economics 3*(1) 601-630.

OECD. (2017). *PISA 2015 Results (Volume III): Students' Well-Being*. Paris: OECD Publishing.

OECD (2020), Gross domestic product (GDP) (indicator). Retrieved October 21 2020 (https://data.oecd.org/gdp/gross-domestic-product-gdp.htm)

Pascoe, Michaela C., Sarah E. Hetrick, and Alexandra G. Parker. (2020). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth 25*(1) 104-112.

Pekkarinen, Tuomas. (2012). Gender Differences in Education. *IZA Discussion Paper No. 6390*.

Putwain, David. (2007). Researching academic stress and anxiety in students: some methodological considerations. *British Educational Research Journal 33*(2) 207-219.

Putwain, David. (2009). Assessment and examination stress in Key Stage 4. *British Educational Research Journal 35*(3) 391-411.

Roberts, Chris, J. Freeman, Oddrun Samdal, . . . 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*(Suppl 2) 140-150.

Schraml, Karin, Aleksander Perski, Giorgio Grossi, and Margareta Simonsson-Sarnecki. (2011). Stress symptoms among adolescents: The role of subjective psychosocial conditions, lifestyle, and self-esteem. *Journal of Adolescence 34*(5) 987-996.

Segool, Natasha K., John S. Carlson, Anisa N. Goforth, Nathan von der Embse, and Justin A. Barterian. (2013). Heightened Test Anxiety among Young Children: Elementary School Students' Anxious Responses to High-Stakes Testing. *Psychology in the Schools 50*(5) 489-499.

Smyth, Emer, and Joanne Banks. (2012). High stakes testing and student perspectives on teaching and learning in the Republic of Ireland. *Educational Assessment, Evaluation and Accountability 24*(4) 283-306.

Sonmark, Kristina, Emmanuelle Godeau, Lily Augustine, Magnus Bygren, and Bitte Modin. (2016). Individual and Contextual Expressions of School Demands and their Relation to Psychosomatic Health a Comparative Study of Students in France and Sweden. *Child Indicators Research 9*(1) 93-109.

Wang, Liang Choon. (2016). The effect of high-stakes testing on suicidal ideation of teenagers with reference-dependent preferences. *Journal of Population Economics 29*(2) 345-364.

MacKinnon, James G. and Matthew D. Webb (2020). When and How to Deal with Clustered Errors in Regression Models. *Queen's Economics Department Working Paper No. 1421*

Van Houtte, Mieke. (2017) Gender Differences in Context: The Impact of Track Position on Study Involvement in Flemish Secondary Education. *Sociology of Education 90*(4) 275-295.

Verger, Antoni, Lluís Parcerisa, and Clara Fontdevila. (2019). The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms. *Educational Review 71*(1) 5-30.

West, Patrick, and Helen Sweeting. (2003). Fifteen, female and stressed: changing patterns of psychological distress over time. *Journal of Child Psychology and Psychiatry 44*(3) 399-411.

Whitney, Camille R., and Christopher A. Candelaria. (2017). The Effects of No Child Left Behind on Children's Socioemotional Outcomes. *AERA Open 3*(3).

Wiklund, Maria, Eva-Britt Malmgren-Olsson, Ann Öhman, Erik Bergström, and Anncristine Fjellman-Wiklund. (2012). Subjective health complaints in older adolescents are related to perceived stress, anxiety and gender – a cross-sectional school study in Northern Sweden. *BMC Public Health 12*(1) 993.

Wing, Coady, Kosali Simon, and Ricardo A. Bello-Gomez. (2018). Designing Difference in Difference Studies: Best Practices for Public Health Policy Research. *Annual Review of Public Health 39*(1) 453-469.

von der Embse, Nathaniel, Dane Jester, Devlina Roy, and James Post. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders 227* 483-493.

Östberg, Viveca, Ylva B. Almquist, Lisa Folkesson, Sara Brolin Låftman, Bitte Modin, and Petra Lindfors. (2015). The Complexity of Stress in Mid-Adolescent Girls and Boys. *Child Indicators Research 8*(2) 403-423.

Appendix

Table S1 – Descriptive statistics and measurements for individual and country level variables

Variable	**Measurement**	**Descriptive statistics**

School-related stress	How pressured do you feel by the schoolwork you have to do?
	(0) Not at all
	(3) A lot
		M=1.26
		SD=0.89

Grade level	What class are you in? Country specific Grade (11 years old) Country specific Grade (13 years old) Country specific Grade (15 years old)	11 years: 31.79 % 13 years: 34.18 % 15 years: 34.03 %

Gender	Are you a boy or a girl?
	(0) Boy
	(1) Girl
		M=0.52
		SD=0.50

Age	Age of respondent, in years.	M=13.59
		SD=1.65

Lives with father	Please answer this question for the home where you live all or most of the time and tick the people who live there.
	(1) Father
		M=1.21
		SD=0.41

Lives with mother	Please answer this question for the home where you live all or most of the time and tick the people who live there.
	(1) Mother
		M=1.05
		SD=0.22

Perception of the economic standing of the family	How well off do you think your family is?
	(5) Very well off
	(1) Not at all well off
		M=2.39
		SD=0.87

Quality of family relations – Father	How easy is it for you to talk to your father about things that really bother you?
	(1) very easy
	(4) very difficult
		M=2.37
		SD=1.14

Quality of family relations - Mother	How easy is it for you to talk to your mother about things that really bother you?
	(1) very easy
	(4) very difficult
		M=1.83
		SD=0.94

Body image	Do you think that you are…?	

	(0) Just right (1) A bit too skinny, or A bit too fat (combined) (2) Much too skinny, or Much too fat (combined)	

Binge drinking	Have you ever had so much alcohol that you were really drunk?
	(1) No, never
	(5) Yes, more than 10 times
		M=1.55
		SD=1.05

Physical activity	Over the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?
	(1) 0 days
	(8) 7 days
		M=3.98
		SD=2.06

Experiences of being bullied	How often have you been bullied at school in the past couple of months?
	(1) I have not been bullied at school the past couple of months
	(5) Several times a week
		M=1.52
		SD=0.99

Physical fighting	During the past 12 months, how many times were you in a physical fight?
	(1) I have not been in a physical fight in the past 12 months
	(5) 4 times or more
		M=1.79
		SD=1.25

Injury	During the past 12 months, how many times were you injured and had to be treated by a doctor or nurse?
	(1) I was not injured in the past 12 months
	(5) 4 times or more
		M=1.78
		SD=1.11

GDP per capita	Yearly national gross domestic product per capita in current prices (1000 US dollars), adjusted for purchasing power.	M=30.77
		SD=11.97

Economic inequality	Yearly national GINI index	M=29.53
		SD=4.05

Youth unemployment rate, %	Yearly national unemployment for population less than 25 years old.	M=19.21
		SD=8.21

Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Table S2 – Descriptive statistics for country level variables. Table shows average values between 2002 and 2010.

Country	GDP per capita, 1000$	Economic inequality, GINI index	Youth unemployment rate, %
Austria	39.43	26.65	9.03
Belgium, Flemish Region	36.84	27.20	20.61
Belgium, Walloon region	38.13	27.01	21.11
Bulgaria	14.38	33.29	21.03
Croatia	25.41	25.08	16.87
Denmark	39.18	25.20	10.79
Estonia	20.47	33.70	19.44
Finland	36.26	25.70	20.47
France	33.67	28.17	25.31
Germany	37.57	28.76	10.43
Greece	26.64	33.82	34.63
Hungary	20.07	27.12	19.75
Iceland	41.50	24.87	11.66
Ireland	44.11	30.99	18.00
Italy	33.02	31.52	28.50
Latvia	17.63	36.65	22.34
Lithuania	15.60	34.28	22.67
Luxembourg	87.02	28.09	17.50
Malta	26.02	28.36	13.73
Netherlands	42.49	26.26	10.32
Norway	52.74	25.95	9.66
Poland	17.50	31.28	31.17
Portugal	25.81	35.55	26.18
Romania	16.20	35.52	22.00
Slovakia	25.02	26.54	30.48
Slovenia	26.51	23.72	16.26
Spain	30.70	33.00	35.29
Sweden	40.94	25.25	22.07
England	35.06	33.10	15.22
Scotland	35.81	32.86	16.07
Wales	35.83	32.88	16.09

Data from Eurostat (2020a; 2020b) and OECD (2020).

Table S3 – Linear regression models with school stress as the outcome. Three way interactions between high-stakes testing, gender, and grade level.

	m1
	b
High-stakes testing	-0.015 (0.030)
Girl	0.066*** (0.013)
Grade level (ref: 11 years)	
13 years	0.176*** (0.011)
15 years	0.227*** (0.016)
High-stakes testing * Girl	0.097* (0.041)
High-stakes testing * 13 years	0.056 (0.032)
High-stakes testing * 15 years	0.083** (0.029)
Intercept	0.585
Country, year, grade fixed effects	Yes
Two-way interactions: country, year, grade	Yes
Demographic covariates	Yes
Additional student covariates	Yes
N individual level	325,176
N country-grade level	93

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Notice that, since the model controls for a full set of two-way interactions between grade level and, respectively, country and year, the main effect of grade level is not very meaningful, as it represent the effect in Austria in 2002. However, the focus here is on the interaction terms, which are not affected by the choice of reference category for country or survey year.

Table S4 – Sensitivity analyses. Linear regression models with school stress as the outcome.

	Placebo test: Lead values of high-stakes testing	Placebo test: Lead values of high-stakes testing	Placebo test: Testing for other purposes	Placebo test: Testing for other purposes	Exclude countries with high grade retention	Exclude high grade retention countries	Dummy coded school stress (linear probability model)	Dummy coded school stress (linear probability model)
	b	b	b	b	b	b	b	b
High-stakes testing	0.027 (0.024)	-0.004 (0.033)	0.038 (0.042)	0.071 (0.046)	0.116*** (0.019)	0.054 (0.027)	0.039*** (0.009)	0.013 (0.014)
Girl	0.080*** (0.013)	0.078*** (0.013)	0.080*** (0.013)	0.087*** (0.014)	0.063*** (0.012)	0.043*** (0.012)	0.034*** (0.006)	0.026*** (0.006)
High-stakes testing * Girl		0.065 (0.042)		-0.064 (0.042)		0.121** (0.040)		0.052* (0.019)
Intercept	0.575	0.576	0.576	0.581	0.555	0.569	0.105	0.110
Country, year, grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-way interactions: country, year, grade	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Additional student covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N individual level	325,176	325,176	325,176	325,176	268,479	268,479	325,176	325,176
N country-grade level	93	93	93	93	78	78	93	93

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010. Tests for other purposes include tests used for “monitoring schools and/or the education system” and tests used for “identifying individual learning needs”). Countries with high grade retention include Belgium (Walloon region), France, Luxembourg, Portugal, Spain.

Table S5 – Linear regression models with school stress as the outcome. Separate models for survey years.

	Year 2002	Year 2002	Year 2006	Year 2006	Year 2010	Year 2010
	b	B	b	b	b	b
High-stakes testing	0.079 (0.059)	0.000 (0.055)	0.073 (0.051)	0.031 (0.044)	0.076 (0.054)	0.036 (0.039)
Girl	0.064** (0.019)	0.051* (0.022)	0.076*** (0.016)	0.063** (0.020)	0.093*** (0.014)	0.082*** (0.016)
High-stakes testing * Girl		0.151* (0.057)		0.081 (0.041)		0.076 (0.048)
Intercept	0.261	0.271	0.210	0.220	0.212	0.220
Country and grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes
Additional student covariates	Yes	Yes	Yes	Yes	Yes	Yes
N individual level	88,534	88,534	122,436	122,436	114,206	114,206
N country-grade level	23	23	29	29	27	27

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Table S6 – Linear regression models with school stress as the outcome. Separate models for grade levels.

	Age 11	Age 11	Age 13	Age 13	Age 15	Age 15
	b	B	b	b	b	b
High-stakes testing	0.068* (0.032)	0.048 (0.040)	0.046 (0.029)	0.077* (0.032)	0.031 (0.071)	0.026 (0.069)
Girl	-0.015 (0.014)	-0.017 (0.015)	0.068*** (0.014)	0.071*** (0.014)	0.189*** (0.019)	0.187*** (0.024)
High-stakes testing * Girl		0.036 (0.043)		-0.060** (0.020)		0.010 (0.044)
Intercept	0.192	0.194	0.373	0.372	0.279	0.280
Country and year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes
Additional student covariates	Yes	Yes	Yes	Yes	Yes	Yes
N individual level	103,372	103,372	111,156	111,156	110,648	110,648
N country-grade level	31	31	31	31	31	31

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Table S7 – Linear regression models with school stress as the outcome. Tests for covariate balance.

	Lives with father	Lives with mother	Economic standing of family	Talk to father	Talk to mother	Body image	Binge drinking	Physical activity	Bullied	Physical fight	Injury
	b	b	b	b	b	b	b	b	b	b	b
High-stakes testing	-0.023 (0.012)	-0.008 (0.008)	-0.035 (0.040)	-0.001 (0.015)	-0.011 (0.013)	-0.009 (0.012)	0.083 (0.095)	-0.042 (0.097)	0.007 (0.034)	-0.105* (0.045)	0.048 (0.035)
Intercept	0.499	0.778	3.630	-0.892	-1.339	2.073	0.251	2.445	0.247	1.438	-0.111
Country, year, grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-way interactions: country, year, grade	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Additional student covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N individual level	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176
N country-grade level	93	93	93	93	93	93	93	93	93	93	93

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Table S8 – Linear regression models with school stress as the outcome. Tests for covariate balance.

	Lives with father	Lives with mother	Economic standing of family	Talk to father	Talk to mother	Body image	Binge drinking	Physical activity	Bullied	Physical fight	Injury
	b	b	b	b	b	b	b	b	b	b	b
High-stakes testing	-0.027 (0.013)	-0.011 (0.008)	-0.026 (0.041)	0.017 (0.021)	-0.000 (0.016)	-0.077** (0.027)	0.117 (0.101)	-0.018 (0.103)	0.019 (0.035)	-0.172 (0.042)	0.039 (0.041)
Girl	-0.024 ** (0.002)	-0.001 (0.001)	0.047 (0.007)	0.370 (0.010)	-0.101 *** (0.008)	0.211 *** (0.013)	0.007 (0.012)	-0.430 (0.028)	-0.059 *** (0.010)	-0.692 *** (0.022)	-0.110 *** (0.010)
High-stakes testing * Girl	0.007 (0.005)	0.005 (0.003)	-0.015 (0.011)	-0.027 (0.028)	-0.023 (0.018)	0.140** (0.045)	-0.065 (0.064)	-0.057 (0.078)	-0.024 (0.014)	0.125* (0.049)	0.013 (0.032)
Intercept	0.515	0.779	3.598	-1.102	-1.270	1.898	0.245	2.691	0.285	1.768	-0.040
Country, year, grade fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Two-way interactions: country, year, grade	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Demographic covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Additional student covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N individual level	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176	325,176
N country-grade level	93	93	93	93	93	93	93	93	93	93	93

Cluster robust standard errors in parantheses. ***p<0.001 **p<0.01 *p<0.05. Data from the Health Behaviour in School-aged Children (HBSC) study, years 2002, 2006 and 2010.

Figure S1 – Histogram showing distribution of school stress

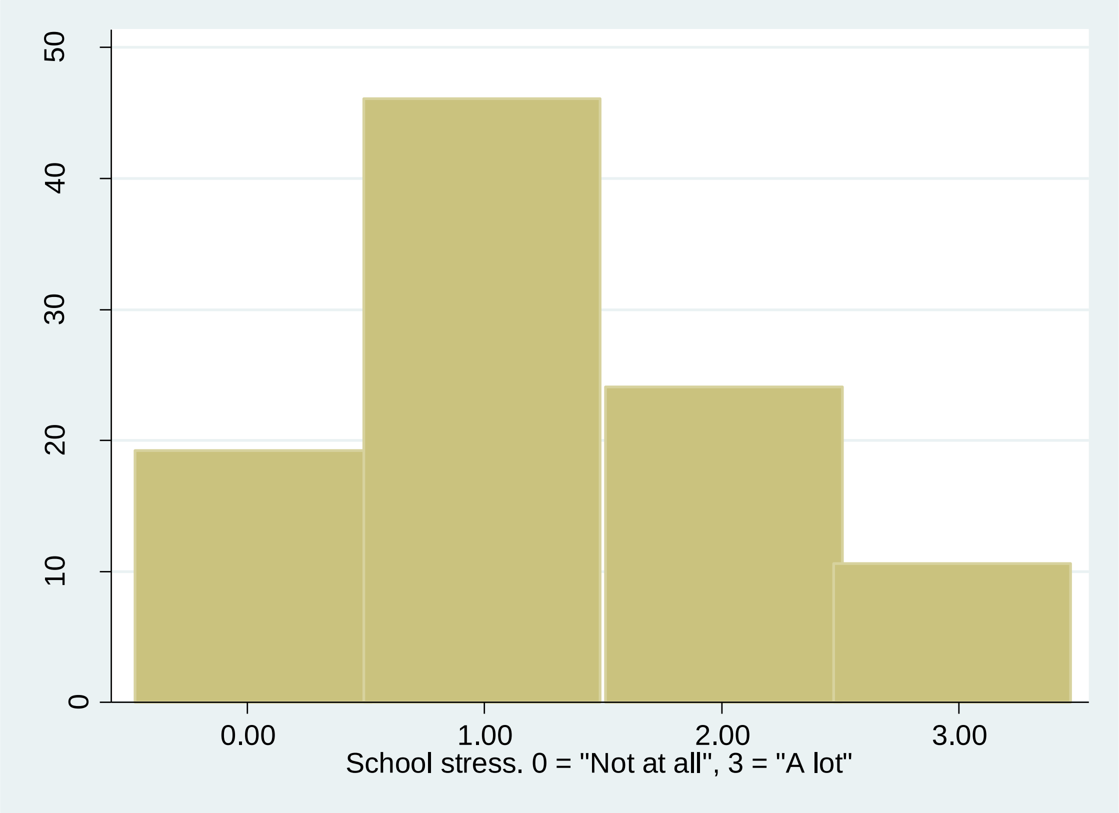
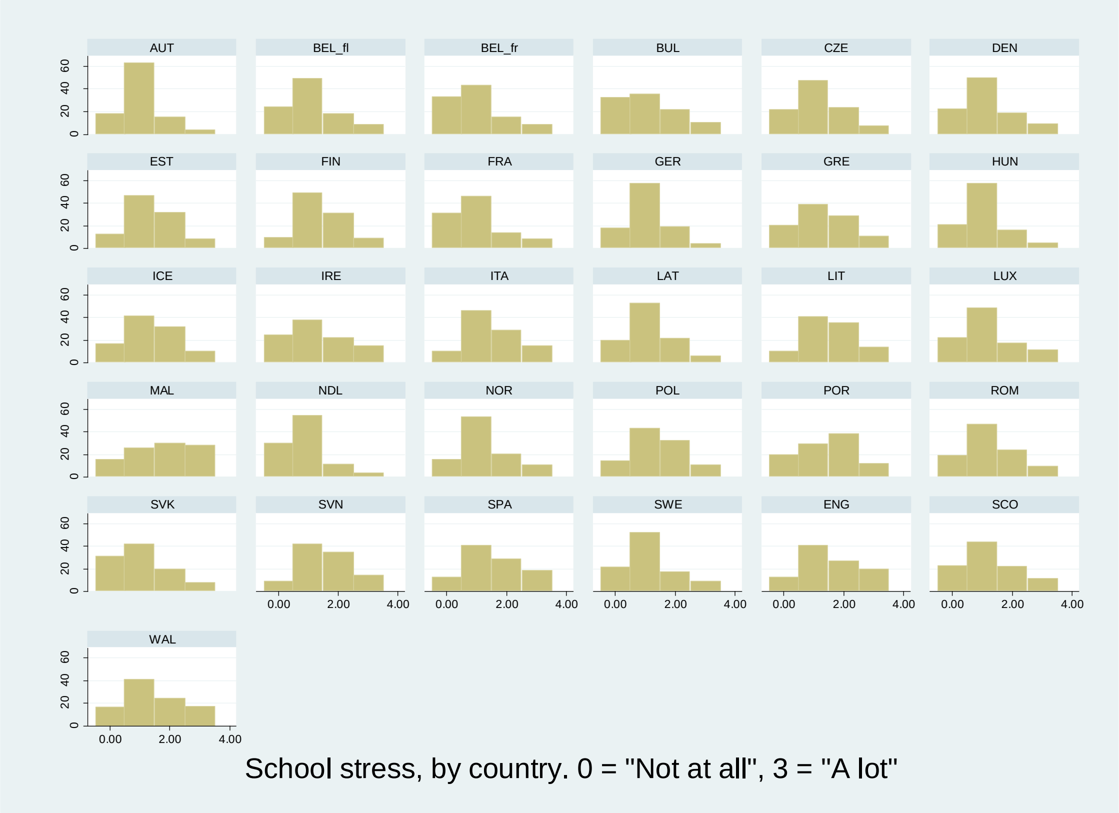


Figure S2 – Distribution of school stress, by country


AUT
BEL_fl
BEL_fr
BUL
CZE
DEN
EST
FIN
FRA
GER
GRE
HUN
ICE
IRE
ITA
LAT
LIT
LUX
MAL
NDL
NOR
POL
POR
ROM
SVK
SVN
SPA
SWE
ENG
SCO
WAL
School stress, by country. 0 = "Not at all", 3 = "A lot"