



Consequences of school grading systems on adolescent health: evidence from a Swedish school reform

Björn Högberg , Joakim Lindgren , Klara Johansson , Mattias Strandh & Solveig Petersen

To cite this article: Björn Högberg , Joakim Lindgren , Klara Johansson , Mattias Strandh & Solveig Petersen (2021) Consequences of school grading systems on adolescent health: evidence from a Swedish school reform, Journal of Education Policy, 36:1, 84-106, DOI: [10.1080/02680939.2019.1686540](https://doi.org/10.1080/02680939.2019.1686540)

To link to this article: <https://doi.org/10.1080/02680939.2019.1686540>



© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 02 Nov 2019.



[Submit your article to this journal](#)



Article views: 4294



[View related articles](#)



[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)



Consequences of school grading systems on adolescent health: evidence from a Swedish school reform

Björn Högberg ^a, Joakim Lindgren ^c, Klara Johansson ^c, Mattias Strandh ^a
and Solveig Petersen ^b

^aDepartment of Social Work, Umeå University, Umeå, Sweden; ^bDepartment of Epidemiology and Global Health, Umeå University, Umeå, Sweden; ^cDepartment of Applied Educational Science, Umeå University, Umeå, Sweden

ABSTRACT

Education reforms that entail increased emphasis on high-stakes testing, assessment and grading have spread across education systems in recent decades. Critics have argued that these policies could have consequences for stress, identity, self-esteem and the overall health of pupils. However, these potentially negative consequences have rarely been investigated in a systematic and rigorous way. In this study we use a major education reform in Sweden, which introduced grades and increased the use of testing for pupils in the 6th and 7th school year (aged 12 to 13 years), to study the consequences of grading and assessment for health outcomes. Using data from the Health Behaviours of School-Aged Children Survey, we find that the reform increased school-related stress and reduced the academic self-esteem of pupils in the 7th school year. This, in turn, had an indirect effect on psychosomatic symptoms and life satisfaction for these pupils. Moreover, the negative effects of the reform were generally stronger for girls, thereby widening the already troubling gender differences in health. We conclude that accountability reforms aimed at increased use of testing, assessment and grading can potentially have negative side effects on pupils' health.

ARTICLE HISTORY

Received 2 April 2019
Accepted 22 October 2019


KEYWORDS

Stress; health; grading; accountability; education reform; gender

Introduction

A global trend towards accountability and assessment in education has characterized education reform across the world in recent decades, leading to increased politicization of education policy (Lingard, Martino, and Rezai-Rashti 2013; Figlio and Loeb 2011). Accountability reforms have entailed a stronger focus on the measurement and quantification of performance through, for example, grading, high-stakes testing or other forms of summative assessment. While the stated aim of accountability reforms has been to raise standards, critics argue that the reforms reproduce social inequalities and are driven by an ideological agenda, with little regard for the health and wellbeing of pupils (Au 2008).

CONTACT Björn Högberg  bjorn.hogberg@umu.se  Department of Social Work, Umeå University, Umeå SE-901 87, Sweden

 Supplemental data for this article can be accessed [here](#)

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The vast majority of quantitative evaluations of accountability reforms have looked at the consequences for academic outcomes, such as test results (Figlio and Loeb 2011). However, accountability reforms have had unintended side effects on teaching practices, educational content and the overall school situation of pupils (Banks and Smyth 2015), and qualitative studies indicate that this, in turn, could have negative health-related consequences for pupils in the form of, for instance, stress (Reay and William 1999; Putwain 2009; see also Gustafsson, Allodi Westling, and Alin Åkerman 2010). However, in their 2015 PISA report, the OECD (2017) claimed that testing frequency was not related to anxiety among pupils and Whitney and Candelaria (2017) found no consistent evidence of negative health effects of school accountability laws in the USA.

To date, few large-scale quantitative studies have examined the health-related consequences of accountability policies, such as grading (Figlio and Loeb 2011; Whitney and Candelaria 2017). Existing evidence is scattered and often circumstantial (West and Sweeting 2003; Sonmark et al. 2016), while explicit tests of policy impacts are rare. Against the background of deteriorating psychosomatic health among adolescents, especially in Northern Europe (Potrebny, Wiium, and Lundegård 2017), as well as a simultaneous increase in school-related stress in a number of countries (Klinger et al. 2015), calls have been made for more research into the role of education policies for the health and wellbeing of pupils (Whitney and Candelaria 2017).

In this paper, we use a recent reform of the Swedish grading system that increased the use of grading, assessment and test-based teaching for pupils in school years 6 and 7 (aged 12–13 years) in order to investigate the importance of systems of assessment for various health-related outcomes. Qualitative evaluations indicate that many pupils perceived stress when grades were introduced (Löfgren and Löfgren 2016). However, thus far, the health effects have not been tested empirically in a rigorous way. Large-scale reforms of grading systems are uncommon, but this Swedish reform provides a compelling quasi-experiment that allows us to investigate the importance of grading systems for pupils' health.

Health is a concept that not only captures the absence of health problems (negative health outcomes) but also the presence of health assets (positive health outcomes). In this study we aimed to capture health effects broadly by assessing both positive and negative health outcomes, in general terms and also specifically related to the school context, as well as capture the somatic and psychological aspects of health. The specific indicators used in the study are further described in the data section.

Background

The Swedish school system and grading reform

Sweden has undergone a transformation from a centralist and state-led education system to a dispersed and marketized system. Public and independently run schools are fully tax funded, but independent schools are mostly for-profit schools (Rönneberg, Lindgren, and Lundahl 2019). In this study, we look at pupils in school years 5, 7 and 9 of compulsory school, when pupils are between 11 and 16 years of age.

Motivated by declining Swedish results in international assessments, especially the OECD Programme for International Student Assessment (PISA) (cf. Pettersson, Prøitz,

and Forsberg 2017), the Swedish centre-right government announced a large-scale education reform in 2009. The centrepiece of the reform was a new national curriculum aimed at strengthening the focus on performance, assessment and goal attainment (see below). Central to the reform was also an extension of formal grading from year 8 (age 14 years) to years 6 and 7 (age 12 and 13 years) and a concomitant increased use of testing to harmonize the grades (Olovsson 2015). The extension of grades was justified by a need to monitor schools and pupils, increase the demands on schools, and provide more information to parents regarding their children's performance in school. Considerations about health-related consequences were largely absent from the political and legislative process that preceded the grading reform. The only reference in the preparatory work was to stress, in which the legislation suggested that early grades could make pupils used to grading, thereby reducing stress (Ds 2010:15). Thus, the architects behind the reform used the issue of stress as an argument in favour of earlier grading.

While accountability reforms in many education systems have focused on standardized high-stakes testing (Lingard, Martino, and Rezai-Rashti 2013), the Swedish public debate has been uniquely focused on grading, largely because teacher-assigned grades are the primary instrument used to sort pupils into different schools and programmes in upper secondary school, and subsequently for tertiary education (Lundahl, Hultén, and Tveit 2017). Until year 9, grades are primarily used to track progress, although many pupils perceive grades as high-stakes already well in advance of year 9 (Swedish National Agency for Education 2017; Låftman, Almquist, and Östberg 2013). Formally, grades in Sweden are primarily high stakes – in the sense that they have important institutionalised consequences – for pupils, not schools. However, in school systems based on competition between schools, such as in Sweden, the grade point average of a school is used to attract pupils to the school, and grades are considered high stakes for teachers as they are used to hold teachers accountable for the performance of their pupils (Lundahl, Hultén, and Tveit 2017; Silfver, Sjöberg, and Bagger 2016). Thus, in a Swedish context, grades, in combination with the extensive use of national standardised tests, can be seen as a functional equivalent to high-stakes testing, implying an emphasis on summative assessment, goal attainment, standardization and monitoring. Accordingly, the Swedish grading reform has been described as intensifying the trend towards an 'outcome-based accountability system' in Swedish schools (Lundahl, Hultén, and Tveit 2017).

The grading reform, implemented in 2012, had two consequences of relevance to this study. Firstly, formal end-of-year grades were extended to pupils in years 6 and 7 who had previously only received informal feedback regarding their performance. Secondly, standardized national tests were moved from year 5 to year 6 as a way of harmonizing the new grades. The national tests were also greatly expanded, from being limited to the core subjects (Swedish, English and maths) to including several additional tests in science and social science.¹ Since the national tests in years 6 and 9 were directly tied to the end-of-year grades, unlike those performed in year 5, they were also given a more 'high-stakes' character, or at least were perceived as such by many pupils (Olovsson 2015). In this study, we look at the first cohort to receive grades and be subject to national tests in year 6, although one year after they first received their grades in 2012.

The extension of grades and national testing to year 6 must be understood in the context of the overall direction of the reform agenda, in particular the new curriculum that was implemented in 2011. As stated, the new curriculum introduced stronger

elements of accountability and assessment for individual pupils, as well as schools, with a focus on goal attainment and measurement of performance (Lundahl, Hultén, and Tveit 2017). The reform thereby intensified the tendency in the Swedish school system to exhaustively use tests and other assessments of pupils (Lundahl, Hultén, and Tveit 2017). Tests are used throughout the school year to harmonize the final grades and since the result of each test can set an upper limit for these grades, the design of the system makes it easy for pupils to fail. Moreover, pupils must achieve at least a 'Pass'-grade in the core subjects in order to be eligible for upper secondary school, which increases school failure rates and probably contributes to higher levels of stress (Gustafsson, Allodi Westling, and Alin Åkerman 2010; Giota and Gustafsson 2017). Thus, unlike test-based systems, school-related pressure in the Swedish system is not concentrated on specific test periods but is continuous throughout the school year, thereby increasing the risk of chronic stress (cf. Hallsten, Josephson, and Torgén 2005; Wheaton et al. 2013).

This performative pressure, which is built into the system as a whole, is experienced by all pupils who receive grades, including pupils in years 8 and 9. The study design rests on the fact that the extension of grades and associated tests to year 6 thereby also extends the performative pressure on pupils who had previously not experienced such pressure, or had experienced it less intensely. While tests were used in years 6 and 7 before the reform, they were not as frequent and were presumably not perceived as high stakes.

The grading reform has been the focus for two qualitative studies. Löfgren and Löfgren (2016) interviewed pupils in year 6, who provided mixed responses, with some experiencing higher levels of motivation and others greater stress when receiving grades. A common theme appears to have been a greater tendency to compare the grades with peers, which could add to the performative pressure in school. Based on an ethnographic study of pupils in year 6, Olovsson (2015) found that many pupils perceived greater performance pressure in relation to the new grades and national tests, but simultaneously showed more discipline and motivation in their schoolwork.

Previous research and theoretical framework

As stated, little empirical research has been conducted in evaluating the role of grading practices on pupils' health. However, there are both empirical and theoretical reasons to expect that increased use of grading, and assessment in general, can be of importance to health. Particularly relevant in this regard is research on high-stakes testing and test anxiety. Whitney and Candelaria (2017) used differences across US states in the implementation of high-stakes testing related to school accountability laws (e.g. No Child Left Behind), but only found evidence of moderate effects on anxiety, and no evidence of effects on sadness.

Other studies have investigated health-related outcomes in relation to specific high-stakes tests or assessments instead of large-scale reforms. West and Sweeting (2003) found that pupils in Scotland perceived more psychosomatic symptoms in proximity to national exams, while Wang (2016), in a study of high-stakes testing in Korea, found similar results but with suicidal ideation as the focal outcome.

Several qualitative and quantitative studies have also shown that (high-stakes) testing is related to feelings of anxiety and stress (von der Embse, Barterian, and Segool 2013; Putwain 2009; Ryan and Ryan 2005; Segool et al. 2013; Banks and Smyth 2015; Silfver,

Sjöberg, and Bagger 2016), as well as to higher cortisol levels, a strong indication of stress (Heissel et al. 2018). Moreover, stress in school is related to health (Sonmark et al. 2016) and Swedish adolescents report that pressure at school is more stressful than pressure at home (Schraml et al. 2011).

In relation to the literature on test anxiety and/or high-stakes testing, it is important to distinguish between tests (or other form of assessments or grades) that are high stakes for pupils, and tests that are high stakes for schools (Banks and Smyth 2015; Whitney and Candelaria 2017). Health-related effects are likely to be stronger when tests or grades have consequences for pupils, like in Sweden, where they are important for progress through the education system.

Theoretically, grading and assessment, through processes of social comparison and social relations, can be expected to be related to self-esteem and self-worth (Ball 2003; Schraml et al. 2011). In the words of Elstad (2010), school is a powerful social institution, and through this institution, society signals to pupils that educational performance is important for social status and esteem. Most forms of assessment in school, particularly formal grading, imply that pupils are categorized and differentiated according to their performance, and are explicitly or implicitly ranked relative to each other (Låftman, Almquist, and Östberg 2013). Thus, the practice of grading sends clear signals to pupils of their place in an officially sanctioned hierarchy, and poor grades can be perceived as a stigma with implications for both identity and self-esteem (Wang 2016; Gustafsson, Allodi Westling, and Alin Åkerman 2010; Putwain 2009; Reay and William 1999). Overall, grading and assessment tend to generate an ‘ethics of competition and performance’ (Ball 2003: 218), in which caring relationships between pupils and teachers are displaced by valuations of pupils in relation to their performance (Ryan and Ryan 2005; Silfver, Sjöberg, and Bagger 2016). This, in turn, intensifies the tendency of constructing one’s self-worth based on external validation, with potentially negative consequences for health (Schraml et al. 2011; Ommundsen, Haugen, and Lund 2005; Ryan and Ryan 2005; Hallsten, Josephson, and Torgén 2005). Assessment, moreover, signals to pupils that how they do in school is decisive for their future prospects in society, and grades can be interpreted by pupils as a sign of whether they will succeed or not when they grow up (Elstad 2010; Låftman, Almquist, and Östberg 2013; Banks and Smyth 2015).

It is not surprising, then, that pupils frequently rank grades and other forms of assessment as one of the most significant stressors in the school environment (Låftman, Almquist, and Östberg 2013; Östberg et al. 2015). Formal grades play a particularly prominent role in this regard. While pupils probably also have a sense of their performance relative to their peers in the absence of grades, the saliency of grades makes performance more explicit, to individual pupils themselves, as well as to their peers (Marsh et al. 2007). Accordingly, research shows that formal grades, compared to tests of cognitive ability, are much more strongly correlated with academic self-concept and self-esteem (Vogl, Schmidt, and Preckel 2018). One interpretation of this is that by making the relative performance of pupils more explicit, grades intensify processes of social comparison and competition (Marsh et al. 2007; see also Wang 2016; Ball 2003; Lundahl, Hultén, and Tveit 2017). Grades, compared to informal or formative assessments, introduce more opportunities for pupils to formally fail, which can negatively affect their self-esteem (Schraml et al. 2011; Giota and Gustafsson 2017; Ryan and Ryan 2005).

There are, moreover, reasons to expect that these effects are stronger for girls than for boys. Firstly, girls tend to place a higher value on, and experience more pressure from, schoolwork, and girls' health is more sensitive to school-related stress (West and Sweeting 2003; Sonmark et al. 2016). Secondly, performance-based self-esteem, amplified by external evaluations, is more common among girls (Hallsten, Josephson, and Torgén 2005; Schraml et al. 2011; Låftman, Almquist, and Östberg 2013). We can therefore expect that an increased and more salient assessment of school performance is perceived as being more stressful for girls.

Based on these considerations, we can formulate five hypotheses in relation to the 2011/2012 Swedish grading reform, which introduced formal grades in years 6 and 7.

H1. School-related stress and low academic self-esteem increased for pupils after grading was introduced.

H2. The effect of grading on school-related stress and academic self-esteem was stronger for girls than for boys.

H3: Psychosomatic symptoms increased and life satisfaction decreased for pupils after grading was introduced.

H4: The effect of grading on psychosomatic symptoms and life satisfaction was stronger for girls than for boys.

H5: The effect of grading on psychosomatic symptoms and life satisfaction can partially be accounted for by school-related stress and low academic self-esteem.

Data and methods

Data

We use individual-level survey data from the Swedish part of the international Health Behaviours of School-aged Children (HBSC) survey. HBSC is a school-based survey that has been conducted every four years since the 1980s, with a focus on the health and health-related behaviours of children and adolescents aged 11 to 15 years. Swedish data were collected by Statistics Sweden on behalf of the Public Health Agency of Sweden, a government body responsible for monitoring and promoting public health in Sweden. Data were collected in 2009/2010 and 2013/2014 (henceforth 2010 and 2014, respectively), using a two-stage cluster design in which a random sample of Swedish schools were drawn at stage one, and then one school class per school year was drawn at random from that school. All pupils in the school class were invited to answer the survey anonymously in the classroom under the supervision of a teacher. The head teachers of each school informed parents about the survey and stated that participation was voluntary. The final sample size was around 7,000 pupils per survey. Response rates at the individual level (the number of schools that declined to participate is not reported) were 88% in 2010 and 69% in 2014, with no systematic differences in trends in response rates between school years (Public Health Agency of Sweden 2014).

The essential advantages of the data, given the aim of this study, are, firstly, that identical questions were asked in both 2010 (before the grading reform) and in 2014 (after the grading reform) and secondly, that these identical questions were asked to pupils in years 5, 7 and 9, respectively. These two features of the data enabled us to employ a differences-in-differences design (see below).

Dependent variables

Hypothesis 1 and 2 refer to school-related stress and low academic self-esteem, respectively. Stress can be conceptualized as a reaction to demands that are perceived as being difficult to manage (Wheaton et al. 2013). Consequently, school-related stress concerns perceived stress directly associated with demands in school. Stress reactions can be assessed using biometric measures, such as cortisol levels, or through self-reported feelings of stress and anxiety, which are in focus in this study. School-related stress was measured using the question: ‘Do you feel stressed by your schoolwork?’, with possible answers ranging from 1 ‘Not at all’ to 4 ‘A lot’.

Regarding academic self-esteem, the HBSC measurement is based on the question ‘In your opinion, what does your class teacher(s) think about your school performance compared to your classmates?’, with possible answers ranging from (1) ‘Below average’ to (4) ‘Very good’. The indicator captures the pupils’ perception of their ability *relative to their classmates*, which, given that academic self-esteem is based on self-evaluation and is therefore sensitive to both external assessment and social comparison (Ommundsen, Haugen, and Lund 2005; Vogl, Schmidt, and Preckel 2018), is important for this study. However, readers should keep in mind that the focus on the view of the teacher makes the indicator somewhat different from most indicators used to capture self-esteem (or self-concept) in empirical research, in which the pupils’ view of their own ability is typically key (Marsh et al. 2007).

Both school-related stress and academic self-esteem are related to other aspects of health (Gustafsson, Allodi Westling, and Alin Åkerman 2010), and both indicators have been previously used in studies of education policy and health outcomes [name deleted to maintain the integrity of the review process].

Hypotheses 3 and 4 refer to psychosomatic symptoms and life satisfaction, respectively. Subjective complaints that are either psychological (e.g. feeling nervous), or somatic (e.g. having a headache), or both, have previously been called psychosomatic symptoms (Potrebny, Wium, and Lundegård 2017). Accordingly, we use this term to describe a combination of psychological and somatic symptoms. We measure psychosomatic symptoms using the HBSC symptoms checklist (HBSC-SCL) (cf. Sonmark et al. 2016). Questions were asked about frequency of headaches, stomach aches, dizziness, backache, sleeping difficulties, feeling low or depressed, being nervous and being irritable or bad tempered, with possible answers ranging from 0 (‘rarely or never’) to 4 (‘about every day’). Based on these eight questions, we generated an additive index, ranging from 0 to 32, with higher values indicating more frequent symptoms and therefore poorer health. Psychosomatic symptoms reflect negative aspects of health (i.e. complaints).

It should be noted that the HBSC-SCL index may capture symptoms related to clinical diagnosis, as well as less serious symptoms. However, in general, these symptoms have

been associated with functional impairment (van Geelen and Hagquist 2016), and interviews with adolescents suggest that the questions generally resonate with the understanding of adolescents (Haugland and Wold 2001).

Life satisfaction reflects positive and evaluative aspects of health and wellbeing. The intention is to capture how each individual balances different aspects of life against each other in order to make an overall appraisal of their current state. Measures of satisfaction with life may reflect social desirability and norms, but have also been shown to reliably predict, for instance, suicide, longevity and other health-related outcomes (Diener, Inglehart, and Tay 2013). In the current study, life satisfaction was measured using the Cantril Ladder, which was accompanied by the following text: 'Here is a picture of a ladder. Suppose the top of the ladder represents the best possible life for you and the bottom of the ladder the worst possible life. Where on the ladder do you feel you personally stand at the present time?' The range is from 0 to 10, with higher values representing a higher level of satisfaction. The Cantril Ladder has shown acceptable reliability and validity in adolescent samples (Levin and Currie 2014).

Independent variables

Our focal independent variables are gender, school year (which, since Swedish classes are homogenous by age, also serves as a proxy for age) and time. Gender is measured as a dummy variable for girls, with boys as the reference category. School year distinguishes between years 5, 7 and 9, with year 7 used as the reference category. Unfortunately, HBSC contains no data on pupils in years 6 or 8. As an indicator of the grading reform, we use time or survey year. Specifically, we enter 2014 as a dummy variable, with 2010 as the reference category. In order to control for compositional differences across the survey years, we enter indicators of parental non-employment (reported by pupils) and socioeconomic status, respectively. Socioeconomic status is measured using the HBSC family affluence scale, which measures the consumption level of the household. Additional information on all variables is provided in Tables S1–S3 in the appendix.

Analytical strategy

A key motivational factor for this study is that the reform can be seen as a policy-induced quasi-experiment. This is because the introduction of grading only affected certain pupils (those in school years 6 and 7), while leaving pupils above or below these years unaffected. These 'untreated' pupils (in years 5 and 9) can therefore be used as a control group. With repeated cross-sectional data and this kind of quasi-experimental setting, difference-in-differences (DID) estimation techniques are applicable. DID techniques use data for the outcome pre and post reform – and the fact that only part of the population was affected – to estimate the differential effect of the reform on those affected by it (the 'treatment group') compared to those not affected (the 'control group') (Imbens and Wooldridge 2009; Whitney and Candelaria 2017). Specifically, we compare the change over time (pre vs. post reform, or 2010 vs. 2014) in the respective outcome variables for the treatment group (year 7), with the change over time in the same outcome variables for the control group (years 5 and 9). We perform the DID analysis using a regression framework, specifically using a series of multilevel linear

regression models. Further technical details of the models are provided in Annex B in the online appendix.

The major benefit of a DID framework compared to a simple pre-post reform comparison for the affected pupils is that the DID framework makes use of a control group, thereby using variation over time, as well as variation between treatment and control groups, thus mimicking an experimental situation. Each age group's (school year's) score on the outcome in 2010 is used as a control, which effectively accounts for all time-invariant unobserved heterogeneity (i.e. unobserved differences across the groups that might bias the results). Thus, time trends that are common to all groups do not cause bias. The school year level variable, in turn, means that we control for differences between treatment (year 7) and control (years 5 and 9) groups that were present before the reform, thus capturing all time-invariant differences between the groups (Imbens and Wooldridge 2009).

One essential assumption required for causal interpretations of the DID estimate is the parallel trends assumption (Lechner 2011). In this setting, the parallel trends assumption amounts to the assumption that in the absence of the grading reform, the difference in the outcomes between the age groups would have remained constant. In other words, the control groups can serve as a counterfactual for the treatment group. If the differences between the age groups would also have changed in the absence of the reform, or in other words, if unobserved heterogeneity is time-varying, the estimates will be biased. The parallel trends assumptions cannot be formally tested since we do not know what would have happened in the absence of the reform (hence 'counterfactual'), although a visual inspection of the time trends for the respective age groups provides some information regarding whether or not the assumption is reasonable. The means of the four outcomes variables (school stress, low academic self-esteem, psychosomatic symptoms and life satisfaction) from 2002 to 2014 are plotted in Figures S1–S4 in the online appendix. Note that the lines do not need to be flat, only that there are no systematic and differential trends between the treatment and control groups. Overall, the average levels of all four outcome variables appear to have varied somewhat across the surveys, but without a clear pattern over time and across school year levels. For three of the four outcome measures, there is a sharp break – an increase (stress and low self-esteem; Figures S1–S2) or decline (life satisfaction; Figure S4) – for pupils in school year 7, with no equivalently sharp break or clear trend for other school years. The exception is psychosomatic symptoms: both years 5 and 7 had an upward trend in symptoms before 2010, while year 9 saw a sharp increase after 2010 (Figure S3). Thus, we might have reason to view the results for psychosomatic symptoms with more caution.

Another assumption is that the composition of the treatment and control groups are not affected by the reform (Lechner 2011). If the sample is representative of each school year level in Sweden, this is a plausible assumption since grade retention is very uncommon in Swedish schools. The sample in HBSA is designed to be representative of the population of Swedish pupils, but this cannot be guaranteed due to selective non-response. However, we include controls for socioeconomic background and parental unemployment to control for potential compositional differences between the cohorts. Overall, the estimates are very similar, whether or not these covariates are included in the models, increasing our confidence that the results are not biased by compositional changes across the surveys.

A third assumption is no spillover effects, that is, that the control groups (pupils in years 5 and 9) are not affected by the reform (Lechner 2011). In this context, spillover effects would be present if pupils in year 5 perceived stress due to anticipation of grades in year 6. To the extent that this is the case, it would lead to a downward bias (i.e. attenuation) of the estimate of the effect of the grading reform for pupils in year 7. Another threat is the fact that standardized national tests were moved from year 5 to year 6 as part of the grading reform. If these tests previously affected pupils in year 5 negatively, this would lead to an upward bias of the estimate of the grading reform.

The new curriculum applies to all pupils, but if the effect of the new curriculum differed between pupils in different school years, this could introduce bias. For example, grades are more high stakes in year 9 as they are used to sort pupils in upper secondary school. Thus, if heterogeneous effects are present, the most probable scenario is that potential effects would be stronger for pupils in year 9 (Banks and Smyth 2015), which would lead to a downward bias of the estimate of the effect of the grading reform in year 7.

Results

We begin with a brief summary of the results in order to make the following section easier to follow. The results presented in Table 1 show that the introduction of grades was associated with increased school-related stress and reduced academic self-esteem for pupils in year 7 (in support of hypothesis 1) and that this increase was roughly equal for girls and boys (contradicting hypothesis 2). The results presented in Table 2 show that psychosomatic symptoms increased, and life satisfaction decreased, for pupils in year 7 after grading was introduced. However, since a similar deterioration was seen for pupils in year 9, this might reflect a general time trend and not the policy *per se* (implying weak support for hypothesis 3). The change in life satisfaction, but not in psychosomatic symptoms, was stronger for girls, implying greater gender gaps for pupils in grade 7 (partially supporting hypothesis 4). Changing levels of stress and academic self-esteem, in turn, accounted for all of the increase in psychosomatic symptoms and around half of the reduction in life satisfaction (in support of hypothesis 5). For the sake of brevity and in order to use consistent statistical terminology, the coefficients will be discussed in terms of 'effects'. We discuss the grounds for making the causal interpretations of the estimates in the method and discussion sections.

Models 1a and 1b in Table 1 together test hypothesis 1, stating that school stress and low academic self-esteem increased after grading was introduced. In model 1a, stress is regressed on school year, time and their interaction, as well as gender and socioeconomic background. The focal coefficients are those for time (with 2010 as the reference category), which show the effect for pupils in year 7 and the interaction terms between time and school year. School stress in year 7 is significantly ($p < 0.001$) higher in 2014 compared to 2010, and the interaction terms show that this positive effect is significantly stronger for pupils in year 7 compared to years 5 and 9. Specifically, the coefficient for time shows that school stress in year 7 was on average 0.27 scale points higher in 2014, but only around 0.13 ($0.27 - 0.14 = 0.13$) scale points higher in year 5, and 0.08 ($0.27 - 0.19 = 0.08$) scale points higher in year 9. The increase in school stress in year 7 (0.27 scale points) is slightly larger than the difference between girls and boys (0.23), not a trivial

Table 1. Multilevel linear regression models with school stress and low academic self-esteem as dependent variables.

Dependent variable	Model 1a	Model 1b	Model 2a	Model 2b
	School stress	Low academic self-esteem	School stress	Low academic self-esteem
Father employed (ref: not employed)	-0.03 (-0.08,0.02)	-0.06 (-0.11,-0.02)	-0.03 (-0.08,0.02)	-0.06 (-0.11,-0.02)
Mother employed (ref: not employed)	-0.02 (-0.06,0.03)	-0.01 (-0.06,0.03)	-0.02 (-0.06,0.02)	-0.01 (-0.05,0.03)
FAS	0.00 (-0.01,0.01)	-0.02 (-0.03,-0.02)	0.00 (-0.01,0.01)	-0.02 (-0.03,-0.02)
Girl (ref: boy)	0.23 (0.20,0.27)	-0.03 (-0.06,0.00)	0.17 (0.10,0.23)	-0.05 (-0.12,0.03)
Time: Year 2014 (ref: 2010)	0.27 (0.21,0.33)	0.17 (0.11,0.22)	0.22 (0.14,0.30)	0.14 (0.06,0.22)
School year (ref = 7): School year 5	-0.28 (-0.34,-0.22)	-0.26 (-0.31,-0.20)	-0.19 (-0.27,-0.11)	-0.25 (-0.32,-0.17)
School year 9	0.54 (0.47,0.61)	0.13 (0.08,0.18)	0.40 (0.32,0.49)	0.11 (0.04,0.19)
Interactions				
Time x School year 5	-0.14 (-0.22,-0.06)	-0.09 (-0.16,-0.01)	-0.12 (-0.23,-0.01)	-0.07 (-0.18,0.03)
Time x School year 9	-0.19 (-0.29,-0.10)	-0.14 (-0.21,-0.06)	-0.19 (-0.30,-0.07)	-0.09 (-0.20,0.02)
Time x Girl			0.09 (-0.01,0.20)	0.06 (-0.05,0.16)
Girl x School year 5			-0.18 (-0.27,-0.09)	-0.02 (-0.12,0.07)
Girl x School year 9			0.26 (0.16,0.37)	0.04 (-0.07,0.15)
Time x Girl x School year 5			-0.05 (-0.18,0.09)	-0.02 (-0.16,0.11)
Time x Girl x School year 9			-0.02 (-0.17,0.13)	-0.08 (-0.23,0.07)
Constant	1.87 (1.79,1.95)	2.46 (2.38,2.53)	1.90 (1.82,1.99)	2.46 (2.38,2.55)
N individuals	13 318	13 230	13 318	13 230
N classes	739	739	739	739
Standard deviation (class level)	0.18	0.12	0.18	0.12
Akaike information criterion	30 987	29 736	30 786	29 742
Log likelihood	-15 481	-14 856	-15 376	-14 854
Likelihood-ratio test			Model 2a vs. 1a Chi2 = 211***	Model 2b vs. 1b Chi2 = 3.76

Individual-level data from HBSC. 95 % confidence intervals in parentheses. Confidence intervals that do not include 0 indicate that estimates are statistically significant at the 5 % level ($p < 0.05$). 'ref' = reference group.

effect, considering the consistent evidence of higher school stress among girls (e.g. Schraml et al. 2011). Model 1b shows that the results are similar with low academic self-esteem as the outcome. Pupils in year 7 had on average around 0.17 points lower academic self-esteem in 2014 compared to 2010, and this change was stronger in year 7 compared to year 5 (where it was 0.08 scale points; $0.17 - 0.9 = 0.08$), and year 9 (where it was 0.03 scale points; $0.17 - 0.14 = 0.03$). The sizes of the effects can be compared by expressing them as standard deviations of the outcome variable, so-called standardised coefficients. These are shown in Table S5 in the appendix. The effect of the reform (the

Table 2. Multilevel linear regression models with psychosomatic symptoms and life satisfaction esteem as dependent variables.

Dependent variable	Model 1a	Model 1b	Model 2a	Model 2b	Model 3a	Model 3b
	Psychosomatic symptoms	Life satisfaction	Psychosomatic symptoms	Life satisfaction	Psychosomatic symptoms	Life satisfaction
Father employed (ref: not employed)	-0.50 (-0.86,-0.14)	0.34 (0.23,0.45)	-0.50 (-0.85,-0.14)	0.34 (0.24,0.45)	-0.36 (-0.69,-0.03)	0.30 (0.20,0.40)
Mother employed (ref: not employed)	-0.53 (-0.85,-0.21)	0.23 (0.14,0.33)	-0.55 (-0.87,-0.23)	0.24 (0.14,0.33)	-0.47 (-0.77,-0.18)	0.22 (0.13,0.31)
FAS (Family affluence)	-0.06 (-0.16,-0.03)	0.09 (0.07,0.11)	-0.05 (-0.12,0.01)	0.09 (0.07,0.11)	-0.05 (-0.11,0.02)	0.08 (0.06,0.10)
Girl (ref: boy)	2.97 (2.76,3.18)	-0.44 (-0.50,-0.38)	2.58 (2.06,3.10)	-0.29 (-0.45,-0.14)	2.35 (2.16,2.55)	-0.33 (-0.39,-0.27)
Time: Year 2014 (ref: 2010)	0.93 (0.47,1.39)	-0.49 (-0.63,-0.35)	0.33 (-0.27,0.93)	-0.23 (-0.41,-0.06)	0.07 (-0.35,0.49)	-0.27 (-0.40,-0.14)
School year (ref = 7):						
School year 5	-1.68 (-2.14,-1.22)	0.55 (0.41,0.68)	-0.74 (-1.34,-0.14)	0.37 (0.19,0.54)	-0.67 (-1.08,-0.25)	0.29 (0.16,0.41)
School year 9	1.19 (0.73,1.66)	-0.52 (-0.65,-0.38)	0.45 (-0.15,1.05)	-0.28 (-0.46,-0.11)	-0.39 (-0.82,0.03)	-0.17 (-0.30,-0.05)
Interactions						
Time x School year 5	-0.47 (-1.12,0.17)	0.27 (0.08,0.46)	-0.34 (-1.18,0.49)	0.09 (-0.15,0.34)	-0.05 (-0.63,0.53)	0.15 (-0.02,0.33)
Time x School year 9	0.25 (-0.39,0.89)	0.13 (-0.06,0.32)	0.47 (-0.36,1.29)	-0.12 (-0.36,0.13)	0.89 (0.31,1.47)	-0.04 (-0.21,0.14)
Time x Girl			1.16 (0.43,1.90)	-0.50 (-0.72,-0.28)		
Girl x School year 5			-1.81 (-2.55,-1.06)	0.35 (0.13,0.57)		
Girl x School year 9			1.47 (0.72,2.21)	-0.46 (-0.68,-0.23)		
Time x Girl x School year 5			-0.30 (-1.34,0.73)	0.34 (0.04,0.65)		
Time x Girl x School year 9			-0.48 (-1.50,0.54)	0.49 (0.19,0.80)		
School stress					2.68 (2.55,2.80)	-0.51 (-0.55,-0.47)
Low academic self-esteem					0.96 (0.83,1.09)	-0.46 (-0.50,-0.42)
Constant	8.65 (8.05,9.26)	6.83 (6.65,7.00)	8.83 (8.19,9.48)	6.76 (6.57,6.95)	1.31 (0.65,1.98)	8.90 (8.70,9.10)
N individuals	12 801	12 771	12 801	12 771	12 801	12 771
N classes	739	739	739	739	739	739
Standard deviation (class level)	1.08	0.31	1.09	0.30	0.93	0.26
Akaike information criterion	82 444	51 225	82 282	51 119	80 422	49 874
Log likelihood	-41 210	-25 600	-41 124	-25 542	-40 197	-24 923
Likelihood-ratio test			Model 2a vs. 1a Chi2 = 172***	Model 2b vs. 1b Chi2 = 117***	Model 3a vs. 1a Chi2 = 2029***	Model 3b vs. 1b Chi2 = 1354***

Individual-level data from HBSC. 95 % confidence intervals in parentheses. Confidence intervals that do not include 0 indicate that estimates are statistically significant at the 5 % level ($p < 0.05$). 'ref' = reference group.

variable 'Time') is slightly stronger with stress (0.31 standard deviations) than with self-esteem (0.22 standard deviations) as the outcome. In sum, hypothesis 1 receives strong support.

The rather strong effects on both stress and self-esteem are contrary to the results of Whitney and Candelaria (2017). One explanation of this is that the NCLB policy in the

USA, unlike the Swedish grades, primarily implies direct consequences for schools, not for pupils (cf. Banks and Smyth 2015).

Hypothesis 2, stating that the effect of grading on school-related stress and self-esteem was stronger for girls, is tested in models 2a and 2b. The focal coefficients are the interaction between time and gender, and the three-way interaction between time, gender and school year. The interaction with gender shows whether the changes in stress and self-esteem differ between girls and boys. For both outcomes, the effect of time is slightly stronger for girls than for boys in year 7 (0.09 scale points for stress and 0.06 scale points for self-esteem) and this stronger effect for girls is slightly weaker in years 5 and 9 (as shown by the negative three-way interaction terms). The direction of the coefficients are in line with the hypothesis, but the interaction terms are not significant, and hypothesis 2 is not supported. The absence of a significant gender difference runs counter to the results of Banks and Smyth (2015), who report more stress among girls in relation to high-stakes tests. However, the Swedish National Agency for Education (2017) reported that, in addition to experiencing stress, many girls also perceived the grades as motivating, in which case the two may have cancelled each other out.

The results of psychosomatic symptoms and life satisfaction as the outcomes are shown in Table 2. Models 1a and 1b test hypothesis 3, stating that symptoms increased and life satisfaction decreased after grading was introduced. The focal coefficients are again those for time and the interaction between time and school year. The coefficients for time are significant in both models and are in the expected direction (positive for psychosomatic symptoms, negative for life satisfaction). Pupils in year 7 had on average almost 1 scale point more psychosomatic symptoms, and half a scale point lower life satisfaction in 2014 compared to 2010. However, the interaction terms show that for neither outcome were these effects significantly stronger in year 7 than year 9. In other words, pupils in year 9 also had more symptoms and lower life satisfaction in 2014, meaning that the results for pupils in year 7 could reflect a broader trend and not the grading reform *per se*. Thus, we find some, though not strong, support for hypothesis 3. Expressed as standardised coefficients, the effect on psychosomatic symptoms corresponds to 0.15 standard deviations and the effect on life satisfaction to -0.25 standard deviations (Table S6).

Hypothesis 4, stating that the effect of grading on psychosomatic symptoms and life satisfaction was stronger for girls than for boys, is tested in models 2a and 2b (Table 2). The interaction between time and gender show that the effects of time on both psychosomatic symptoms and life satisfaction were clearly stronger for girls than for boys. In fact, the main effect of time in model 2a, which shows the change in symptoms for boys in year 7, is weakly positive (though not significant), meaning that all the increase in symptoms in year 7 is driven by the increase for girls. The three-way interaction terms between time, gender and school year are in the expected direction (negative), but not significant, with psychosomatic symptoms as the outcome (model 2a). With life satisfaction as the outcome, however, the negative effect for girls was clearly stronger in year 7 than years 5 or 9, as shown by the significant three-way interaction terms (model 2b).

To facilitate interpretation of the results, we present the three-way interactions graphically in Figures 1 and 2. The figures show predicted levels of psychosomatic symptoms and life satisfaction (vertical axes) for boys (grey lines) and girls (black lines) in years 5 (dotted lines), 7 (solid lines) and 9 (dashed lines), in both 2010 (left-

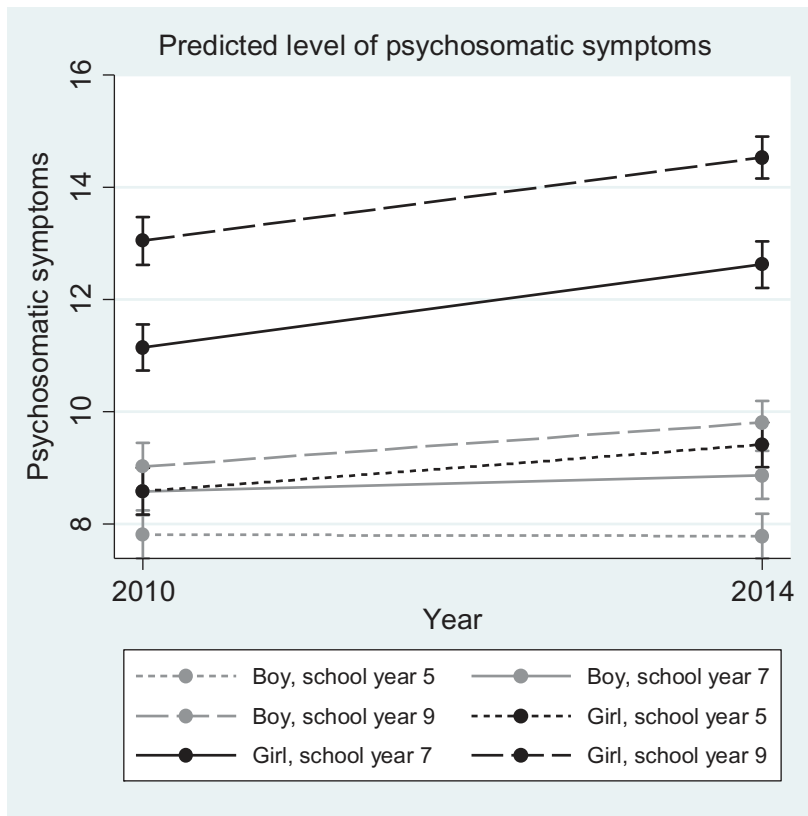


Figure 1. Predicted level of psychosomatic symptoms.

Higher values indicate more symptoms.

hand side) and 2014 (right-hand side), respectively. The three-way interaction is assessed by comparing how the *change* in the gender gap between 2010 and 2014 differs across the school years. In other words, we compare the gaps between girls and boys within the respective school years on the left-hand side of the figures with the corresponding gaps on the right-hand side of the figures. Looking first at [Figure 2](#), with life satisfaction on the vertical axis, boys and girl in year 5 had nearly equal levels of life satisfaction in both 2010 and 2014, and although we see a slightly larger decline for girls, the gender gap remains fairly stable. In year 9, girls clearly had lower life satisfaction in both periods, but the gender gap is identical. However, the gender gap in year 7 more than doubles between 2010 and 2014, from around 0.3 to almost 0.8 scale points. In other words, while the gender gap in life satisfaction was mostly stable for pupils in years 5 and 9, it more than doubled for pupils in year 7 after the reform.

A similar pattern of a larger gender gap for pupils in year 7 can be seen in [Figure 1](#), with psychosomatic symptoms on the vertical axis. In 2010, the difference was around 2.5 scale points and in 2014 more than 3.7 scale points. However, in this instance, the gender gap also grew larger in years 5 and 9, although less clearly so. We therefore find strong support for hypothesis 4 with regard to life satisfaction, but less so for psychosomatic symptoms.

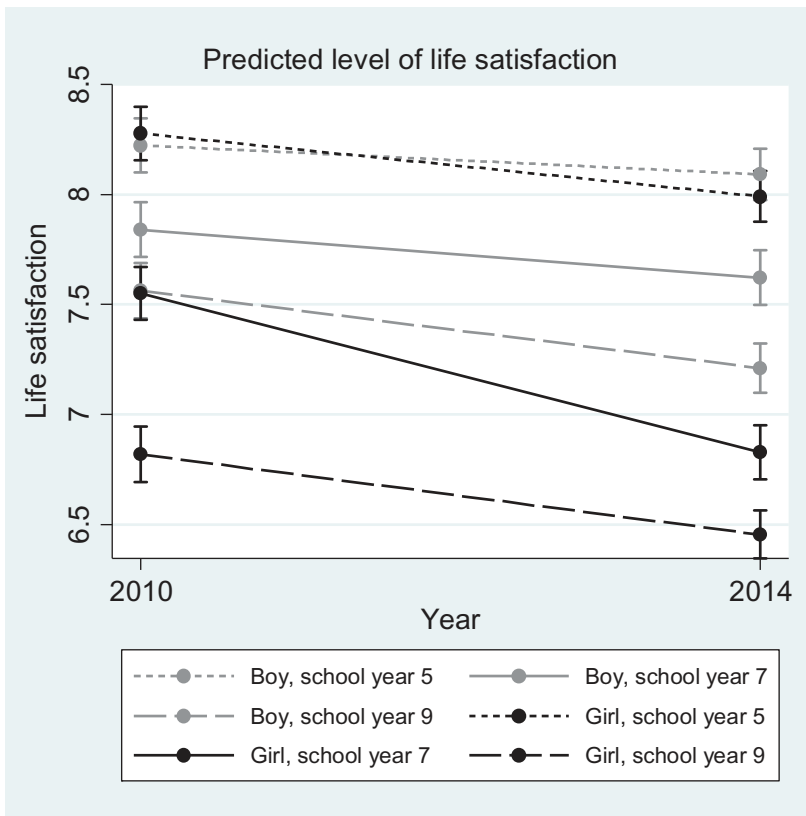


Figure 2. Predicted level of life satisfaction.
Higher values indicate higher life satisfaction.

Models 3a and 3b (in Table 2) test hypothesis 5, stating that the effect of grading on psychosomatic symptoms and life satisfaction can partly be accounted for by school stress and academic self-esteem. The extent to which this hypothesis is supported can be assessed by comparing the coefficients for time in models 1a and 1b with the equivalent coefficients in models 3a and 3b, in which we introduce stress and self-esteem as mediators. Thus, what we call an indirect effect refers to the effect of the reform that is ‘transmitted’ by school stress and academic self-esteem. If there was an effect of the reform on psychosomatic symptoms and life satisfaction, it is likely that this was partially due to how the reform affected stress and self-esteem. Then we would expect to see that the change in psychosomatic symptoms and life satisfaction is smaller if we hold the level of stress and self-esteem constant over the two time points, as we do by controlling for stress and self-esteem in models 3a and 3b.

The strong effect of time (year 2014) on psychosomatic symptoms (0.93, model 1a) disappears when school stress and academic self-esteem are held constant in model 3a (where the coefficient is 0.07). Thus, the effect on psychosomatic symptoms is completely accounted for by school stress and lower academic self-esteem, meaning that if pupils in year 7 had not perceived more stress and lower self-esteem in 2014, their level of symptoms would have been stable. The equivalent indirect effect on life satisfaction is

slightly weaker: the coefficient is reduced from -0.49 (model 1b) to -0.27 (model 3b), meaning that around half of the reduction in life satisfaction can be accounted for by increased stress and lower self-esteem. A further notable result from model 3a is that the interaction term for year 9 increases to 0.89 scale points when we hold stress and self-esteem constant. Thus, although pupils in year 7 and year 9 saw a roughly equal increase in psychosomatic symptoms between 2010 and 2014, the increase for grade 9 cannot be accounted for by increased stress and lower self-esteem.

The significance of the indirect effects are formally tested using structural equation models, with bootstrap methods used to calculate standard errors and confidence intervals. Detailed results are shown in Table S4 in the appendix. With both psychosomatic symptoms and life satisfaction, the indirect effects of school stress and academic self-esteem were significant. We therefore find support for hypothesis 5.

Sensitivity analyses

We have performed several sensitivity analyses to probe the robustness of the results and conclusions. Firstly, we have dichotomized school stress and academic self-esteem (as described in Table S7 in the appendix) and re-estimated the models in Table 1, but using logistic regression. The results were substantially similar to those in Table 1, with the exception that the increase in school stress was not significantly stronger in year 7 compared to year 5.

Secondly, we have re-estimated the models in Table 2, but using an indicator of self-rated general health as the outcome. The results, shown in Table S8 in the appendix, are substantially similar to those presented in Table 2, with partial support for hypotheses 3 and 4 and strong support for hypothesis 5.

Thirdly, we have performed a number of 'placebo tests' to rule out alternative explanations, for example, that the estimates presented here only pick up general but unobserved time trends. Specifically, we have estimated the same models but using the equivalent HBS data from Norway and Denmark, countries similar to Sweden in terms of culture (e.g. individualism and secularism) and institutions (e.g. welfare state regimes). If similar results were found in these countries, it would cast doubt on the proposed cause of the Swedish results (i.e. the grading reform). Note that the logic behind this approach is equivalent to a differences-in-differences-in-differences approach. However, in this case, a formal differences-in-differences-in-differences approach would require running interaction models with multiple four-way interaction terms, as well as mediation analyses of three-way interaction terms, making the results unwieldy. The results (available on request) showed that all hypotheses were clearly rejected in both Norway and Denmark. In sum, the health trajectories of Swedish year 7 pupils stand out in a Scandinavian comparison.

Discussion

This study examined a major accountability reform of the Swedish school system; a reform centred on the introduction of grades and increased use of testing, especially standardised national tests, in the 6th and 7th school year. Specifically, the study investigated the effects of the reform on, firstly, school-related stress and academic self-esteem,

and secondly, psychosomatic symptoms and life satisfaction, as well as potential gender differences in these effects. We found evidence that the reform increased stress and reduced academic self-esteem among pupils, and some evidence that it also led to more psychosomatic symptoms and reduced life satisfaction which, in turn, was largely accounted for by the increased stress and reduced self-esteem. In the case of life satisfaction, the negative effect of the reform was stronger for girls, thus increasing the gender gap.

Overall, the results were weaker with psychosomatic symptoms as the outcome, while the most robust conclusions can be drawn with regard to the directly school-related measures – stress and self-esteem – and life satisfaction. One interpretation of this is that increased focus on assessment mainly affects more proximate (i.e. directly school-related) or evaluative and ‘positive’ (i.e. life satisfaction) outcomes, but that the consequences are not strong enough to directly impact on somatic or psychological health symptoms. Psychosomatic symptoms, as measured in HBSC, could capture many symptoms that are not related to the school environment (e.g. menstrual pain). Thus, the indicator might contain ‘noise’, making the estimates less precise. A second interpretation is that any potential effect of the reform on psychosomatic symptoms was offset by other factors. The fact that the increase in symptoms in year 7 was fully accounted for by changes in school stress and academic self-esteem (Table 2, model 3a), while the corresponding increase in year 9 appeared to have had other causes, indicates that the deterioration in year 7 could have been at least partially due to the reform. However, this effect was hidden because the ‘control group’ (year 9) was affected by other, unmeasured changes. Moreover, qualitative studies (Löfgren and Löfgren 2016; Olovsson 2015) have found that the responses of year 6 pupils to receiving their first grades were mixed, with some reporting feelings of stress and anxiety, but others reporting that they felt more motivated, or in some cases both simultaneously. Thus, greater stress could have been partially cancelled out by greater motivation, resulting in a smaller increase in symptoms.

Overall, the results of this study are in line with qualitative studies that have reported negative health consequences of accountability policies such as testing and grading (Reay and William 1999; Putwain 2009; Låftman, Almquist, and Östberg 2013; Silfver, Sjöberg, and Bagger 2016), as well as with some quantitative evidence that points towards similar conclusions (West and Sweeting 2003; Wang 2016; Sonmark et al. 2016). Not least, the generally stronger effects on girls compared to boys are in line with studies suggesting that girls are more sensitive to performance-based self-esteem and that the health of girls is more sensitive to demands in school (West and Sweeting 2003; Schraml et al. 2011; Låftman, Almquist, and Östberg 2013; Sonmark et al. 2016). However, other studies have found either no or inconsistent health-related effects of high-stakes testing (Whitney and Candelaria 2017) and testing frequency (OECD 2017). The divergence between the results could be because Whitney and Candelaria (2017) investigated tests that were primarily high stakes for schools, not pupils.² This indicates that studies of health effects of test-based accountability policies should also focus on policies that directly increase the demands faced by pupils.

It is notable that the Swedish grading reform was explicitly motivated by the declining Swedish results in the PISA study (Ds 2010:15; see also Pettersson, Prøitz, and Forsberg 2017). A recent investigation of cross-national trends in school-related stress from the early 1990s until 2010 suggested that differential trends in stress across countries appear

to have been linked to performance in PISA, and to the public debates and school reforms that were triggered by the perceived inadequate PISA results (Klinger et al. 2015). From this perspective, the 'PISA effect' (Grek 2009) may extend beyond the direct effect on education policies, and involve indirect repercussions on pupils' health and overall wellbeing in school.

As stated in the introduction, in recent decades there has been a growing emphasis on testing, assessment and grading in education policy. However, the health-related consequences of these policies have only received scant attention, especially among quantitative researchers. The results of this study suggest that the way in which assessment systems are designed could have important repercussions for the health and overall wellbeing of pupils, including the extent of gender-based inequalities in health. If this is the case, then reforms involving increased testing and grading may need to consider the potential negative side-effects on health and give more weight to the non-academic consequences of the policies when considering how assessments are implemented. This conclusion is in line with recent recommendations to take wellbeing into account when designing and evaluating education policy (OECD 2017). Against this background, it is notable that health-related aspects were barely touched upon in the preparatory work for the Swedish grading reform (Ds 2010:15). It should also be noted that the additional national tests in science and social science introduced in year 6 were removed by the newly-elected centre-left government in 2016, partly motivated by concerns about stress among pupils (SOU 2016).

The Swedish grading reform and Swedish education policy in general are clearly situated in the context of an increased focus on accountability. However, while the results have thus far been discussed in relation to the broader literature on testing and assessment, the Swedish education system contains some idiosyncrasies that must be highlighted in order to make the results comprehensible. Most notable is the emphasis on summative assessments in the form of final grades, assigned by teachers and given by end of the year (Lundahl, Hultén, and Tveit 2017). While accountability policies are often associated with test-based approaches, and the increased use of testing was integral to Swedish grading reform, Swedish education policy has been uniquely focused on grades.

Furthermore, when grades or high-stakes tests become more salient, they tend to permeate most aspects of teaching (e.g. 'teaching to the test'). Thus, in this context, the introduction of grades in years 6 and 7 implied that performative pressure was increasingly present also in situations in which the final grades were not directly in focus (Löfgren and Löfgren 2016; Silfver, Sjöberg, and Bagger 2016). A consequence of this combination of high-stakes final grades and intensive testing throughout the school year is that pressure to perform is constant for pupils, and is not concentrated on specific test periods. Considering the negative consequences of chronic stress (Hallsten, Josephson, and Torgén 2005), such a system could be potentially harmful.

Thus, the results might not be generalizable to other systems of assessment and accountability. Nevertheless, the results could shed light on the secular trend towards more psychosomatic symptoms and mental health problems among adolescents in many countries, including Sweden, an increase that several authors have ascribed to changes in schooling (Potrebny, Wium, and Lundegård 2017; Låftman, Almquist, and Östberg 2013; Gustafsson, Allodi Westling, and Alin Åkerman 2010).

Limitations

The conclusions of this study are conditional on the assumptions behind differences-in-differences analysis being fulfilled. The assumptions and their applicability have already been discussed, but it should be stressed that some assumptions are potentially or partially invalidated in this context. The study could not account for time-varying unobserved confounding, such as time trends specific to year 7. Moreover, the grading reform was not always as ‘neat’ as would be ideal, and pupils in years 5 and 9 also experienced changes associated with the new curriculum. Thus, the treatment and control groups are not perfectly separated. The grading reform was part of a broader reform agenda, and it can be difficult to separate the effects of the grading reform from other aspects of this agenda. However, the other elements of the reform besides the introduction of grades in year 6 and 7 – such as the new curricula – were comparable for all school years included in the analysis, and should therefore be cancelled out by the difference-in-difference approach. Moreover, the consistency of the results across a range of sensitivity tests, and the null results generated by the placebo tests in other Scandinavian countries, provide some credibility to the causal interpretation of the estimates. Nevertheless, we should again emphasise that we do not regard grades *per se* to be the sole or main factor behind the results. Rather, it is the combination of the final end-of-year grades, which are high stakes for pupils, with schooling that is oriented towards intensive and frequent testing throughout the school year.

It should also be emphasised that the situation for pupils in year 6 in 2012/2013 had some exceptional features. In addition to the new grades, year 6 pupils were also subject to an extensive amount of new national tests (many of which were removed in 2016), and there was significant public attention and controversy surrounding the reform at the time (Olovsson 2015). We studied the first cohort to be given grades in years 6 and 7, and the attention given to the introduction of grades in 2012 could have made the experience more dramatic for pupils. On the other hand, we studied the cohort the year after they first received grades, and one year after the national tests, and some habituation could have taken place during this time.

Notes

1. The additional national tests in science and social science were later removed in 2016, after complaints that the administrative burden associated with the tests was too large and that the tests caused stress among pupils (SOU 2016).
2. The OECD (2017) does not provide information regarding which numbers it base its conclusions on, making a comparison with its results difficult.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Forskningsrådet om Hälsa, Arbetsliv och Välfärd [2015-00048]; Vetenskapsrådet [2018-03870_3].

Notes on contributors

Björn Högberg is a researcher at the Department of Social Work, Umeå University. The present study is part of a research project aimed at investigating the consequences of reforms of education systems for the health and wellbeing of pupils. Orcid: 0000-0002-0199-0435.

Joakim Lindgren is an Associate Professor at the Department of Applied Educational Science, Umeå University. His scholarly interests are education policy, evaluation, school inspection and problems of socialisation and juridification in education. Orcid: 0000-0003-2167-6299.

Klara Johansson is a researcher in public health and epidemiology. Her main research interests are adolescent health (mental health, sexual health, and safety), social determinants of health, gender and health, and emerging global health challenges.

Mattias Strandh is a professor at the Department for Social Work, Umeå University, and at the Centre for Research on Child and Adolescent Mental Health, Karlstad University. His research interests include the interrelationship between mental health and school outcomes, as well as the micro-level impact of policy and policy configurations on this interrelationship. Orcid: 0000-0002-6867-6205.

Solveig Petersen has a PhD in Paediatrics and holds an Associate Professorship in Epidemiology and Public Health at Umeå University. Her research focus is the intersection between health and education during childhood. Orcid: 0000-0001-6720-2430.

ORCID

Björn Högberg  <http://orcid.org/0000-0002-0199-0435>

Joakim Lindgren  <http://orcid.org/0000-0003-2167-6299>

Klara Johansson  <http://orcid.org/0000-0002-3749-998X>

Mattias Strandh  <http://orcid.org/0000-0002-6867-6205>

Solveig Petersen  <http://orcid.org/0000-0001-6720-2430>

References

- Au, W. 2008. "Devising Inequality: A Bernsteinian Analysis of High-stakes Testing and Social Reproduction in Education." *British Journal of Sociology of Education* 29 (6): 639–651. doi:10.1080/01425690802423312.
- Ball, S. J. 2003. "The Teacher's Soul and the Terrors of Performativity." *Journal of Education Policy* 18 (2): 215–228. doi:10.1080/0268093022000043065.
- Banks, J., and E. Smyth. 2015. "'your Whole Life Depends on It': Academic Stress and High-stakes Testing in Ireland." *Journal of Youth Studies* 18 (5): 598–616. doi:10.1080/13676261.2014.992317.
- Diener, E., R. Inglehart, and L. Tay. 2013. "Theory and Validity of Life Satisfaction Scales." *Social Indicators Research* 112 (3): 497–527. doi:10.1007/s11205-012-0076-y.
- Ds 2010:15. 2010. *Betyg från årskurs 6 i grundskolan [Grades in school year 6 in compulsory school]*. Departementspromemoria.
- Elstad, J. I. 2010. "Indirect Health-related Selection or Social Causation? Interpreting the Educational Differences in Adolescent Health Behaviours." *Social Theory & Health* 8: 134–150. doi:10.1057/sth.2009.26.
- Figlio, D., and S. Loeb. 2011. "School Accountability." In *Handbooks in Economics*, edited by E. Hanushek, S. Machin, and L. Woessman, 383–421. Vol. 3. The Netherlands: North-Holland.
- Giota, J., and J.-E. Gustafsson. 2017. "Perceived Demands of Schooling, Stress and Mental Health: Changes from Grade 6 to Grade 9 as a Function of Gender and Cognitive Ability." *Stress and Health* 33: 253–266. doi:10.1002/smi.v33.3.

- Grek, S. 2009. "Governing by Numbers: The PISA 'effect' in Europe." *Journal of Education Policy* 24 (1): 23–37. doi:10.1080/02680930802412669.
- Gustafsson, J. E., M. Allodi Westling, and B. Alin Åkerman. 2010. *School, Learning and Mental Health A Systematic Review*. Stockholm: Royal Swedish Academy of Sciences.
- Hallsten, L., M. Josephson, and M. Torgén. 2005. *Performance-based Self-esteem: A Driving Force in Burnout Processes and Its Assessment*. Stockholm: Arbetslivsinstitutet [National Institute for Working Life].
- Haugland, S., and B. Wold. 2001. "Subjective Health Complaints in Adolescence – Reliability and Validity of Survey Methods." *Journal of Adolescence* 24: 611–624. doi:10.1006/jado.2000.0393.
- Heissel, J. A., E. K. Adam, J. L. Doleac, D. N. Figlio, and J. Meer. 2018. "Testing, Stress, and Performance: How Students Respond Physiologically to High-Stakes Testing." NBER Working Paper No. 25305.
- Imbens, G. W., and J. M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86. doi:10.1257/jel.47.1.5.
- Klinger, D.A., J. G. Freeman, S. S. Sebok, L. Bilz, K. Liiv, D. Ramelow, and O Samdal., Dür, W., Rasmussen, M. 2015. "Cross-national Trends in Perceived School Pressure by Gender and Age from 1994 to 2010." *European Journal of Public Health* 25(suppl_2): 51–56
- Låftman, S. B., Y. B. Almquist, and V. 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. doi:10.1080/13676261.2013.780126.
- Lechner, M. 2011. "The Estimation of Causal Effects by Difference-in-Difference Methods." *Foundations and Trends® in Econometrics* 4 (3): 165–224. doi:10.1561/08000000014.
- Levin, K. A., and C. Currie. 2014. "Reliability and Validity of an Adapted Version of the Cantril Ladder for Use with Adolescent Samples." *Social Indicators Research* 119 (2): 1047–1063. doi:10.1007/s11205-013-0507-4.
- Lingard, B., W. Martino, and G. Rezaei-Rashti. 2013. "Testing Regimes, Accountabilities and Education Policy: Commensurate Global and National Developments." *Journal of Education Policy* 28 (5): 539–556. doi:10.1080/02680939.2013.820042.
- Löfgren, R., and H. Löfgren. 2016. *Att få sina första betyg: En rapport om elevers berättelser om sina erfarenheter av att få betyg i årskurs 6 [Receiving grades for the first time: A report on pupils experiences of grades in year 6]*. Stockholm: Swedish National Agency for Education.
- Lundahl, C., M. Hultén, and S. Tveit. 2017. "The Power of Teacher-assigned Grades in Outcome-based Education." *Nordic Journal of Studies in Educational Policy* 3 (1): 56–66. doi:10.1080/20020317.2017.1317229.
- Marsh, H. W., U. Trautwein, O. Lüdtke, J. Baumert, and O. Köller. 2007. "The Big-Fish-Little-Pond Effect: Persistent Negative Effects of Selective High Schools on Self-Concept after Graduation." *American Educational Research Journal* 44: 631–669. doi:10.3102/0002831207306728.
- OECD. 2017. *PISA 2015 Results (Volume III): Students' Well-Being*. Paris: PISA, OECD Publishing.
- Olovsson, T. G. 2015. *Det kontrollera(n)de Klassrummet: Bedömningsprocessen I Svensk Grundskolepraktik I Relation till Införandet Av Nationella skolreformer [The Assessment Process in Swedish Compulsory School Practice in Relation to the Introduction of National School reforms]*. Akademiska avhandlingar vid Pedagogiska institutionen, Umeå universitet, Umeå Studies in the Educational Sciences.
- Ommundsen, Y., R. Haugen, and T. Lund. 2005. "Academic Self-concept, Implicit Theories of Ability, and Self-regulation Strategies." *Scandinavian Journal of Educational Research* 49 (5): 461–474. doi:10.1080/00313830500267838.
- Östberg, V., Y. B. Almquist, L. Folkesson, S. Brodin Låftman, B. Modin, and P. Lindfors. 2015. "The Complexity of Stress in Mid-Adolescent Girls and Boys." *Child Indicators Research* 8: 403–423. doi:10.1007/s12187-014-9245-7.
- Pettersson, D., T. S. Prøitz, and E. Forsberg. 2017. "From Role Models to Nations in Need of Advice: Norway and Sweden under the OECD's Magnifying Glass." *Journal of Education Policy* 32 (6): 721–744. doi:10.1080/02680939.2017.1301557.

- Potrebny, T., N. Wiium, and M. M.-I. Lundegård. 2017. "Temporal Trends in Adolescents' Self-reported Psychosomatic Health Complaints from 1980-2016: A Systematic Review and Meta-analysis." *PloS One* 12 (11): 1–24. doi:10.1371/journal.pone.0188374.
- Public Health Agency of Sweden. 2014. *Skolbarns hälsovanor i Sverige 2013/14 Grundrapport [Health behaviours in school-aged children in Sweden 2013/14 – Report]*. Stockholm: Folkhälsomyndigheten [Public Health Agency of Sweden].
- Putwain, D. W. 2009. "Assessment and Examination Stress in Key Stage 4." *British Educational Research Journal* 35 (3): 391–411. doi:10.1080/01411920802044404.
- Reay, D., and D. William. 1999. "'i'll Be a Nothing': Structure, Agency and the Construction of Identity through Assessment." *British Educational Research Journal* 25 (3): 343–354. doi:10.1080/0141192990250305.
- Rönnerberg, L., J. Lindgren, and L. Lundahl. 2019. "Education Governance in Times of Marketization." In *Handbuch Educational Governance Theorien*, edited by R. Langer and T. Brüsemeister, Educational Governance, Vol. 43, 711–727. Wiesbaden: Springer VS.
- Ryan, K. E., and A. M. Ryan. 2005. "Psychological Processes Underlying Stereotype Threat and Standardized Math Test Performance." *Educational Psychologist* 40 (1): 53–63. doi:10.1207/s15326985ep4001_4.
- Schraml, K., A. Perski, G. Grossi, and M. Simonsson-Sarnecki. 2011. "Stress Symptoms among Adolescents: The Role of Subjective Psychosocial Conditions, Lifestyle, and Self-esteem." *Journal of Adolescence* 34: 987–996. doi:10.1016/j.adolescence.2010.11.010.
- Segool, N. K., J. S. Carlson, A. N. Goforth, N. von der Embse, and J. 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. doi:10.1002/pits.2013.50.issue-5.
- Silfver, E., G. Sjöberg, and A. Bagger. 2016. "An 'appropriate' Test Taker: The Everyday Classroom during the National Testing Period in School Year Three in Sweden." *Ethnography and Education* 11 (3): 237–252. doi:10.1080/17457823.2015.1085323.
- Sonmark, K., E. Godeau, L. Augustine, M. Bygren, and B. 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: 93–109. doi:10.1007/s12187-015-9299-1.
- SOU. 2016. *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning [Equivalent, Objective and Effective – A New National System for assessment]*. 25, Stockholm: Statens offentliga utredning [Official Inquiries of the Swedish Government].
- Swedish National Agency for Education. 2017. *Utvärdering av betyg från årskurs 6 [Evaluation of Grading in School Year 6]*. Report 451. Stockholm: Skolverket [Swedish National Agency for Education].
- van Geelen, S. M., and C. Hagquist. 2016. "Are the Time Trends in Adolescent Psychosomatic Problems Related to Functional Impairment in Daily Life? A 23-Year Study among 20,000 15–16 Year Olds in Sweden." *Journal of Psychosomatic Research* 87: 50–56. doi:10.1016/j.jpsychores.2016.06.003.
- Vogl, K., I. Schmidt, and F. Preckel. 2018. "The Role of Academic Ability Indicators in Big-fish-little-pond Effect Research: A Comparison Study." *The Journal of Educational Research* 111: 429–438. doi:10.1080/00220671.2017.1291485.
- von der Embse, N., J. Barterian, and N. Segool. 2013. "Test Anxiety Interventions for Children and Adolescents: A Systematic Review of Treatment Studies from 2000–2010." *Psychology in the Schools* 50 (1): 57–71. doi:10.1002/pits.2013.50.issue-1.
- Wang, L. C. 2016. "The Effect of High-stakes Testing on Suicidal Ideation of Teenagers with Reference-dependent Preferences." *Journal of Population Economics* 29: 345–364. doi:10.1007/s00148-015-0575-7.
- West, P., and H. Sweeting. 2003. "Fifteen, Female and Stressed: Changing Patterns of Psychological Distress over Time." *Journal of Child Psychology and Psychiatry* 44: 399–411. doi:10.1111/jcpp.2003.44.issue-3.

- Wheaton, B., M. Young, S. Montazer, and K Stuart-Lahman. 2013. "Social Stress in The Twenty-first Century" in Aneshensel, C.S., Phelan, J.C., Bierman, A. (eds.), *Handbook of the Sociology of Mental Health* (pp. 299–323). 2nd ed. Dordrecht: Springer
- Whitney, C. R., and C. A. Candelaria. 2017. "The Effects of No Child Left behind on Children's Socioemotional Outcomes." *AERA Open* 3 (3). doi:[10.1177/2332858417726324](https://doi.org/10.1177/2332858417726324).