The impact of standardized testing on teacher burnout*

Helena Svaleryd (Uppsala University) and Jonas Vlachos (Stockholm University)

Abstract

This paper examines the impact of standardized testing on the mental health of teachers in Sweden. Standardized testing provides a clear set of standards and information on student achievement, but concerns have been raised that it may contribute to stress and burnout. We present quasi-experimental evidence on the impact of reforms to the Swedish system of standardized testing on register-based measures of mental healthcare use, sick leave, and turnover among teachers. The results suggest that the reforms, if anything, resulted in improved mental health and to the very least, substantive negative effects on health can be ruled out. Previous work has focused on the impact of testing and test-based accountability on teacher mobility and sorting, and this study contributes by directly analyzing teacher burnout.

Introduction

Teachers face diverse and often conflicting demands from various stakeholders and as students' needs are boundless, teachers regularly confront morally challenging dilemmas (Berlak and Berlak, 1981). While this is to some extent unavoidable, it may cause stress and burnout. Survey evidence indeed suggests that while teachers report a relatively high degree of job satisfaction, they suffer from higher levels of stress, fatigue, anxiety, and sleep disorders than other professionals (Iriarte Redín et al, 2020). Work-related stress of this kind is regarded as an important risk-factor for depression and related mental health problems, and teacher burnout is a major concern across countries (von der Embse, 2019; EASHW, 2022). Burnout not only affects teachers but also have implications for teacher turnover and absenteeism, as well as for the long-run supply of teachers, plausibly to the detriment of student achievement (Miller et al, 2008; Ronfeldt et al, 2013). While many factors could account for stress in the teaching profession, standardized testing and the corresponding focus on student achievement on these tests are often discussed as contributing factors, and there is some correlational evidence to support this (von der Embse, 2017; Jerrim and Sims, 2022). This suggests a potentially important trade-off between teacher burnout and the benefits from testing, but there is a lack of causal evidence addressing the question. This paper presents quasi-experimental evidence on the impact of standardized testing in Sweden on teacher burnout, using register-based measures of mental healthcare use and sick leave.

Standardized testing has become increasingly common around the world and Sweden is no exception to this trend (Bergbauer et al, 2021; Sievertsen, 2022). From 2009 and 2013, standardized national

* This work is supported by a grant from the Swedish Research Council. Svaleryd is at the Department of Economics at Uppsala University and Vlachos is at the Department of Economics at Stockholm University. Contact: helena.svaleryd@nek.uu.se and jonas.vlachos@su.se.

tests were introduced in lower- and upper-elementary schools, and the number of subjects tested in lower-secondary schools increased from three to ten. Standardized testing offers a set of standards and provides information on how students perform in relation to them. It is not obvious that this should increase stress among teachers. On the one hand, testing increases the workload and can increase pressures to deliver results on outcomes for which teachers are only partially in control. Conflicts can also arise between what teachers regard as the purpose of teaching and an emphasis on test results. On the other hand, testing may help define and limit teachers' professional obligations, thereby reducing the workload and stress. Ultimately, whether or not standardized testing increases teacher stress is an empirical question.

The impact of standardized testing is analyzed using a series of difference-in-differences settings. The main analysis focuses on the 2012 introduction of national tests in grade 6. Due to idiosyncratic difference in schools' grade-configurations there are schools that run from grade 1 to 6, while others only run up to grade 5, leaving their upper-elementary teachers unexposed to the tests. These teachers constitute a close to ideal control group for similar teachers at schools also serving students in grade 6. For the 2010 introduction of tests in grade 3, pre-school teachers act as the control group for the exposed lower-elementary teachers. In the final year (9th grade) of lower-secondary school, tests in natural science subjects were introduced in 2009 and in social science subjects in 2013. In this setting, teachers in always-tested core subjects and those in never-tested practical/aesthetic subjects act as control groups for exposed teachers. As there are some shortcomings in group-comparability in the latter settings, the focus is on the grade 6-reform and the other reforms serve as additional evidence.

None of the education reforms studied showed evidence of harming teachers mental health. In fact, when tests were introduced in grade 6, direct measures of mental health, based on diagnoses from hospitals and prescriptions of antidepressants and tranquilizers, fell by around one percentage point (close to 10 percent). For sick leave due to mental health issues, the estimates suggest a similar percentage point decrease. Further, the estimates suggest null-effects for indicators of teacher turnover. A more detailed analysis shows that the positive impact on health is concentrated to teachers at schools serving students with a relatively strong socioeconomic background. For the grade 3 and grade 9 reforms, there are indications of improved mental health outcomes and to the very least, substantive detrimental effects can be ruled out. However, the estimates less precise and the assumption of parallel trends is more questionable in these settings.

While school-level results on Swedish national tests are publicly available, they are not tied to any national accountability measures. Nevertheless, the tests have real implications for teachers and students and have a substantive impact on teaching. For example, the fraction of teachers who claim that the tests have an important impact on their planning increased from 48 to 73 percent between 2003 and 2015 (Skolverket, 2016).[1] Testing is also a likely contributing factor behind the increased use of assessments to judge teacher effectiveness. Between 2000 and 2018, the fraction of Swedish lower-secondary schools doing so increased from 10 to 40 percent (Sievertsen, 2022). How such school-level assessments are used and what implications they have for individual teachers is unknown but as Figlio and Loeb (2011, p 389) stress, the evidence strongly suggests that information pressure influences educator behavior, also in the absence of formal accountability measures.

[1] The surveys by Skolverket (2016) also show that the fraction of teachers who say that the tests have a substantive impact on the subject-matter covered increased from 33 to 60 percent and the fraction of teachers who did not have students work on test-related problems dropped from 45 to 12 percent.

This paper contributes to the quasi-experimental literature on the impact of testing and test-based accountability on teachers and teacher labor markets by analyzing how testing affect teacher mental health. Previous research has largely focused on the impact on teacher mobility and sorting across schools. Boyd et al (2008) study the introduction of state-mandated testing in New York. As in our setting, the tests had no direct consequences for teachers but still generated considerable pressures to perform. The results show a decrease teacher turnover in the school grade of testing and that more experienced teacher moved into the tested grade relative to other grades. These patterns are relatively uniform across schools serving different types of students suggesting that teachers perceived testing – or the changes it brought about – as beneficial. Also in a setting without direct consequences for teachers, Gjefsen and Gunnes (2020) find that the introduction and publication of a value-added type measure of school performance in Oslo (Norway) triggered a substantial move out from the teaching profession. However, schools were able to fill vacated positions with teachers of similar quality which speaks against a deterioration of the working climate.

Turning to settings with formal accountability measures, Clotfelter et al (2004) find that North Carolina's school-based accountability system adversely affected teacher retention at schools serving low-performing students. Similarly, Feng et al (2018) find substantial increases in teacher mobility away from schools in Florida receiving the lowest accountability grade. For other thresholds in the accountability scheme they find no impact on mobility. To the opposite, Dizon-Ross (2020) finds that schools receiving a low accountability grade in New York City experienced decreased teacher turnover and increased quality of incoming teachers, measured by their value-added. She offers suggestive evidence that this might be due to increased effort by principals. In one of the few papers studying other outcomes than mobility, Reback et al (2014) find that accountability pressures from the US No Child Left Behind Act (NCLB) led to an increase in teachers' concerns for job security and in the fraction of teachers claiming to plan for early retirement. While this is indicative of increased stress and a worsened workplace environment, the impact on teacher health is not studied.

Concerns that testing and accountability may increase pressure on teachers and in other ways have negative consequences for teachers' work situation and health have been raised for a long time (Barksdale-Ladd and Thomas, 2000; Lund and Turner, 2001). The results presented question the commonly held view that testing in itself contribute to a lower quality of the workplace environment and teacher burnout. This, however, is in a setting where standardized tests are used as information provider and how the findings generalize to settings where with higher stakes for teachers is.

1. Theoretical background

Standardized testing offers a clear set of standards and provides information on how students achieve in relation to them. Testing highlights achievements on the tested domains by making them comparable and works as a tool for the alignment of the actual curriculum across schools and teachers. Standardized testing, therefore, affects teachers' work conditions, regardless if explicit incentives are tied to them (Figlio and Loeb, 2011). The most straightforward reason why this may increase stress and burnout is that the time devoted to testing is generally not compensated by reducing the regular workload.

On top of the mechanical increase in the workload comes pressure to deliver strong test results. Teachers may also experience testing as an intrusion on their professional autonomy. Since tests are

limited in scope and domain, this intrusion can generate a perceived need to compromise with professional ideals. As formally discussed by Besley and Ghatak (2018) such conflicts can be framed as a multitasking problem (Holmström and Milgrom, 1991) where teachers have prosocial motivations. While Besley and Ghatak use their framework to analyze incentive pay, the mere act of measuring a limited set of outcomes or tasks can generate discrepancies between ideals and practice if teachers expect such measurement to affect them. Besley and Ghatak further argue that the correspondence between individuals' motivations and the incentives in place affect the self-selection to occupations and workplaces. Introducing testing may therefore generate a particularly sharp conflict between what existing teachers regard as the purpose of teaching and a new emphasis on test results. It is also possible that teachers internalize the failures and stress that testing introduces among students.[2]

The potential strains from testing can be interpreted through widely used public health models of mental strain at the workplace. Following the demand-control-support (Karasek, 1979; Johnson and Hall, 1988), testing can increase stress and burnout by adversely affecting the balance between professional demands and actual control. Similarly, the effort-reward-imbalance model (Siegrist et al, 1986) suggests that testing can increase the gap between effort and reward. As results on standardized tests are strongly related to student characteristics and other external factors, teachers only have partial control of the measured outcomes. The imbalances that testing may contribute to are therefore potentially heterogeneous across settings. The conflict between ideals and demands discussed by Besley and Ghatak (2018) can also be framed as a type of ethical stress, i.e. the physical and psychological response to being constrained from undertaking what is perceived to be the morally right course of action (Ulrich et al, 2007). Such stress can arise if a focus on testing contradicts teachers' perception of students' needs and the purpose of education. There are numerous examples of strategic behavior and questionable practices that arise in response to testing, ranging from outright cheating, to an excessive attention to students at the margin of failure at the expense of others, and "teaching-to-the-test" (Figlio and Loeb, 2011; Sievertsen, 2022). A perceived need to engage in such practices and to shift priorities towards what is tested are clearly potential sources of ethical stress.

These public health models also indicate why increased stress is by no means a necessary outcome of testing. Students' needs are boundless and teachers necessarily face morally challenging dilemmas (Berlak and Berlak, 1981). By narrowing the objectives, testing can help resolve these dilemmas and define professional obligations. Having clearly communicable results may then reduce rather than increase the gap between demands and resource limitations, as well as between effort and reward. Similarly, while testing can certainly contribute to some types of ethical stress it can provide backing for otherwise unpopular courses of action that teachers perceive as being right, for example with respect to students in need of additional resources.

In conclusion, it is not possible to formulate precise theoretical predictions regarding the impact of testing on teacher burnout. On the one hand, standardized testing may contribute to heightened levels of stress by increasing perceived accountability, reducing autonomy in determining teaching content, increasing workload associated with test administration, and pressure caused by students feeling

[2] The impact of testing on student stress and wellbeing is not the subject of this study but there is a growing literature on the subject. For example, Heissel et al (2021) find that US students' stress levels increase when subject to high-stakes tests. For Sweden, Högberg et al (2019) find that the reforms in Swedish upper-elementary school studied in this paper may have had a negative impact on students' psychological wellbeing and life-satisfaction. However, their approach does not allow them to control for general time-trends. On the other hand, Green et al (2022) study the introduction of low-stakes tests in Norwegian lower- and upper elementary schools. They lack direct measures of student stress and wellbeing, but survey responses weakly indicate a positive impact on students' perception of the school climate.

stressed. On the other hand, nationwide standardized test may clarify what material should be taught thereby reducing stress and the workload. Clearly, the impact of testing may differ depending on the composition of students at a given school with schools with more academically advantaged students potentially being less challenging for teachers. Moreover, individual characteristics, such as experience or age, may affect their reaction to testing. Additionally, how school use the test scores for internal accountability and how prospective students and their parents consider them when selecting schools could also influence outcomes. It is, for example, possible that privately run schools may place greater emphasis on accountability concerns due to the profit motive.

2. Institutional setting and reforms

This section provides a brief overview of how the Swedish school system is organized, a description of the reforms to the system of national testing, and changes to the classification of teachers. The purpose is not to give a comprehensive overview of these matters but to highlight the aspects that are relevant to the empirical analysis.

2.1. The organization of the Swedish school system

Swedish compulsory schooling is divided into three stages: lower-elementary (grades 1-3, age 7-9), upper-elementary (grades 4-6, age 10-12), and lower-secondary (grades 7-9, age 13-15). All schools follow the same curriculum and there is no tracking. The division into stages is not binding for schools or teachers and there are schools with grade configurations that do not match the stages. In parts of the analysis, such idiosyncrasies in schools' grade configurations are used for the purpose of empirical identification.

In Sweden, there is free school choice in the sense that all schools, both publicly or privately run, have to admit applicants if slots are available. If publicly managed schools are oversubscribed, priority is based on proximity. Privately run schools following the Swedish curriculum are not allowed to charge fees or tuition and all approved schools are fully funded through vouchers. When oversubscribed, voucher schools mainly rely on waiting lists to determine priority. Some schools admit students based on their talents for music, sports, or the arts but, with a few tightly regulated exceptions, admission based on academic talent or achievement is not allowed. In contrast, admission to non-compulsory upper-secondary schools is almost completely based on academic achievement, measured by the grade point average from the 9th and final grade of compulsory school. Since sorting is not based on achievement until students leave compulsory school, achievement is essentially low-stakes until students reach the lower-secondary stage.

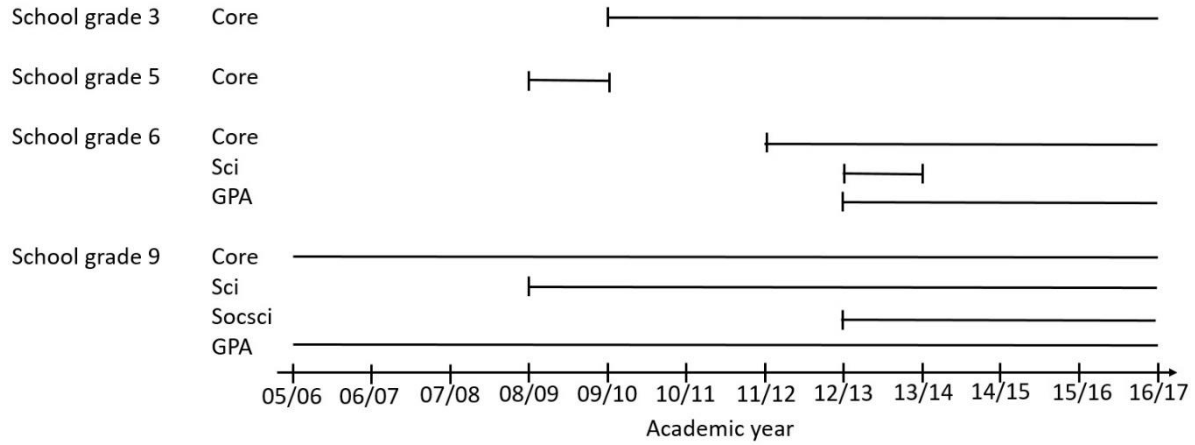
2.2. The system of national testing: purpose and reforms

In the 9th grade, standardized national tests in the core subjects Swedish, English, and mathematics have existed since long prior to the period of analysis in this study. Due to concerns for lower student achievement and varying standards across schools, the testing regime expanded rapidly between 2009 and 2013. In the 2009/10 academic year, tests in Swedish and mathematics were introduced in grade 3, the final year of the lower-elementary stage. These tests consist of a number of subtests that are all on a pass/fail basis. In the academic year 2008/09, tests in core subjects were introduced in grade 5 but they were discontinued in 2009/10. Instead, in 2011/12 tests were introduced in grade 6, the final year of the upper-elementary stage. The following year, 2012/13, tests in the sciences were added in

grade 6, but these tests were discontinued already in 2013/14. In the 9th and final year of lower-secondary school, tests in the sciences (biology, chemistry, and physics) were introduced in 2008/09 and in the social sciences (civics, geography, history, and religion) in 2012/13. Each year, schools are randomly allocated one of the tests in the sciences and one in the social sciences. The subjects that a school is subject to are announced approximately one month prior to the respective test date. The purpose of the random draw and late announcement is to reduce the scope for schools to narrow the curriculum to the subjects they have been assigned. The reforms are summarized in Figure 1.

Figure 1. Time line of implementation and removal of national tests and grading

Note



Note: Reform dates to the system of national testing and grading. Core stands for English (grades 5 and above), mathematics, and Swedish. Sci stands for science subjects (biology, chemistry, and physics). Socsci for social science subjects (civics, geography, history, and religion). GPA means that students receive subject grades.

One of the purposes of the national tests is to aid teachers align subject-grades across schools, although the extent to which they achieve this goal is questionable (Vlachos, 2019). Such subject-grades were introduced in the 6th grade in 2012/13 with students previously only being evaluated using subject-grades during the final stage of compulsory school. In addition, the purpose of the tests is to facilitate comparisons of student learning and achievement. Individual-level test results available for internal evaluations while school-level test results are publicly available online for external evaluations. Since there are no national accountability schemes connected to the tests, they can be considered as low-stakes. How and if the national tests are used by individual schools to hold teachers accountable for student achievement is not well known, but according to the 2018 PISA survey approximately 40 percent of lower-secondary schools judge teacher effectiveness based on student achievement (Sievertsen, 2022). This is an increase from around 10 percent in 2000. Schools are obliged to design a plan of action If a student is expected to fail a course and considerable effort is put into this. Ultimately, teachers decide if students should pass or fail a course and the test results are used as input in this process. Since the tests mainly take place during the final semester of the respective stage, there is limited time to undertake additional efforts and the test results are best seen

as an indication of the success or failure of these efforts. However, schools usually undertake continued and intensified remedial measures when students fail a test.

Over the course of a few years, there were thus a number of changes all increasing the focus on student results. Despite formally being low-stakes, the national tests are seen as important and have a substantive impact on teaching. Surveys show that the fraction of teachers who claim that the tests have an important impact on their planning increased from 48 to 73 percent between 2003 and 2015 (Skolverket, 2016). The fraction of teachers who say that the tests have a substantial impact on the subject-matter covered in class increased from 33 to 60 percent and the fraction of teachers who did not have students work on test-related problems dropped from 45 to 12 percent. Students also regard the tests as important and there are indications that the reforms aiming for a stronger focus on achievement contributed to stress and reduced students' academic self-esteem (Högberg et al, 2021).

2.3. Sick leave and healthcare in Sweden

The present study aims at assessing mental health using register data on use of mental health care services use and sick leave. These measures imply mental health problems with real consequences for both the individual and the schools. Before describing the empirical strategy and data in detail, it may be helpful to understand how the Swedish social insurance system and healthcare system works and the consequences for the coverage of the measures across groups and time.

The social insurance system provides financial security to workers not able to work due to illness. The first two weeks are paid by the employer, thereafter sickness benefits are paid by the public social insurance. After 7 days of absence from work, the illness needs to be confirmed by a doctor. The benefits paid out are based on previous earnings and are capped at a level that implies most teachers in our sample receive 80 percent of their income; in 2023 the maximum sick leave benefit was 80 percent of 43 750 SEK (about 4000 euros) per month. During the studied period there have been some changes to the insurance system, notable changes to the rules, and enforcement of rules, concerning the assessment of work ability. For these and other reasons, the aggregate level of sick leave has varied over the study period, which is important to be aware of in the empirical analysis.

The Swedish healthcare system is a publicly funded system that provides universal coverage to all residents. Basic health and medical care is provided by healthcare centers and more advanced care is provided by specialized care centers and hospitals. The patient can get, but does not need, a referral from primary care, to receive specialized care. User fees are low, about 20 euros per visit, and capped at about 130 euros per year. Because public health care is managed by the Regions (20 in total) the organization of the health care and fees vary somewhat across regions. As for drugs, most prescribed drugs, including antidepressants, sedatives and other drugs prescribed for mental health issues, are heavily subsidized and there is a cap of maximum 230 euros per year. Importantly for this study, financial concerns should not affect the usage of psychiatric care and drugs.

3. Empirical strategy and data

As outlined in the previous section, a series of reforms to the system of testing were implemented between the academic years 2008/09 and 2012/13. This generates a setting suitable for a difference-in-differences approach, provided that appropriate control groups of unexposed teachers exist. For

each of these reforms, some groups of teachers were exposed to a new set of national tests while others were unaffected. We estimate the model (1) using linear regression:

(1) $$y_{ikt} = \gamma_k + \delta_t + \sum_{\tau=-4,\tau\neq-1}^{5} \beta_\tau D_{k\tau} + \boldsymbol{X}_{i,\tau=-1} \times \tau\delta_t + \varepsilon_{it}$$

where y_{ikt} represents the outcome variable for individual *i*, in group *k* in time period t . The model includes fixed effects for group k γ_k and time δ_t. The estimates of β_τ are the differences between the treated teachers and the control group for each time period, compared to the reference year prior to the reform. In some specifications, we include a vector of control variables ***X*** of variables such as age, sex and family characteristics at the time period prior to the reform, interacted with time fixed effects to account for any compositional differences between the treated and control group. To account for correlation within municipalities, standard errors are clustered at the municipality level. To capture the potential lagged impact of the national tests, we follow both groups for up to five years after the reform was implemented.

Additionally, we will conduct a heterogeneity analysis to investigate whether the effects of standardized testing on mental health vary based on factors such as teacher age, student composition at the school and whether the school is privately or publicly run.

As a large fraction of teachers were exposed to changes in the testing system at some point in time, a close-to-ideal control group can only be identified for the introduction of national tests in grade 6. For this reason, most of the analysis is based on upper-elementary teachers while the remaining reforms are used to provide auxiliary evidence. The treated and control groups for each of the reforms are presented below.

3.1. Data

To analyze the impact of the accountability reforms on teachers, we first identify all individuals working as teachers using the Teacher register, which includes the universe of teachers working at schools in grades 1-12. This registry contains information on working hours, work experience, position and school identifier. Via the School register we add information about the schools' grade configuration. Based on the information on school grade configuration and the position of the teacher we identify our treatment and control groups. The population consists of teachers belonging to the treatment and control group in the base year. For the analysis of the implementation of national tests in grade 3 the control group consists of preschool teachers identified by occupation. Information on individual characteristics is collected from the Longitudinal Integration Database for Health Insurance and Labour Market Studies, which integrates data from the population, tax, and social insurance registers. The registry contains information on, for example age, sex, number and ages of children in the household educational attainment, workplace, occupation, income and household identifier making it possible to add information on the partner.

To measure mental health, we use diagnoses from the National Patient register and drug prescriptions from the Drug registry, both held by the National Health Board. We study whether the individual received a diagnosis for mental health disorder (ICD 10 code F) either in in-patient or out-patient care or a drug prescription for an antidepressant, tranquilizers and sleep remedies (ATC codes N05, N06A). The patient register does not contain information on primary care but the Drug registry include all

prescribed and collected drugs in Sweden, also drugs prescribed in the primary care. Information about sick pay for sick periods longer than two weeks is gathered from The National Insurance Board MiDAS database. We study whether the likelihood of a new sick leave spell changed after the reform. Of particular interest is sick leave due to mental health issues. However, since changes in diagnoses practices for mental health may have changed over time, we also study new sick leave spells for any reason. Health outcomes are measured September – August to align with the school year. Other than mental health and sick leave we study whether the reform had any effects on teachers changing workplaces and the probability to continue working as a teacher.

For the heterogeneity analysis we construct a measure of the student composition at each school by predicting the school grade point average (GPA) from a regression of family background variables (e.g. parental education level, income and migration background) on GPA in grade 9. Students are linked to their parents via the Multigenerational register and students are linked to the schools via the School register.

Using the teacher registry, we can identify all teachers working at all schools in Sweden. Assigning teacher to specific classes or school years is however less straight forward. Teachers are not permanently assigned to a specific stage but can move between them. Prior to a teacher certification reform in 2013, teacher mobility between educational stages and subjects was supported by flexible regulation. The scope for such mobility was to some extent reduced by the reform that specified teacher eligibility requirements but the regulation cannot be regarded as particularly strict (Frostenson, 2015).[3] This reform also induced changes to the classification of teachers in the official registers. Until the academic year 2012/13, some teachers were registered as being assigned to a specific stage but a large number of teachers were registered as being assigned to, for example, grades 1-7 or grades 4-7, without further details provided. From 2012/13, the classification became more detailed regarding which specific stages and subjects to which a teacher was assigned. Such changes to the classification of teachers have implications for the empirical analysis as they render it close to impossible to consistently track the entire group of, say, lower-elementary teachers over time. It is however possible to classify parts of the teacher workforce based on their assignment in a specific year and track these individual teachers over time.

3.2. National test in grade 6

Due to idiosyncrasies in the grade configuration, some Swedish schools cover grades 1-5, while others cover grades 1-6. Both types of schools employ upper-elementary teachers, but only those at 1-6-schools are affected by the introduction of the national testing in grade 6 during the academic year 2011/12. In other aspects, teachers at both types of schools have comparable obligations and serve similar types of students. One caveat is that national tests in grade 5 were administered during the 2008/09 and 2009/10 academic years, affecting teachers in both types of schools. A further issue is that teacher set subject-grades were introduced in grade 6 in 2012/13, the academic year following the introduction of national testing. This resulted in teachers at 1-6-schools were exposed to two types of reforms with different content, both of which emphasized measurable student achievement.

[3] Schools can hire teachers that are not formally eligible but they are mildly restricted in what tasks the teachers can perform. Eligibility rules may, however, still affect the allocation of teachers.

The national tests introduced in 2011/12 covered the core subjects of Swedish, English, and mathematics. In 2012/13, additional tests in science and social science subjects were introduced, but these were discontinued in 2014/15. The primary components of these tests take place in March and April during the spring term. However, all three tests have an oral component administered in November and December. The tests are graded along the same scale as subject-grades, i.e. A-F where F stands for a "fail" grade. While a failing the national test does not prevent students from continuing to the lower-secondary stage, it serves as a signal to students and schools that additional effort and resources may be necessary.

3.3. National tests in grade 3

National tests for students in grade 3 in Swedish and mathematics were introduced in 2009/10, consisting of eight sub-tests in Swedish and seven in mathematics. The tests are on a pass/fail basis and serve as a signal that student achievement is lagging and that additional attention is needed. Student progression to the next stage is, however, not normally restricted by failing one or more of the tests. Due to the introduction and expansion of testing in other school stages, there is no unaffected group of compulsory school teachers to use as a control group. Instead, we rely on preschool teachers as the counterfactual to lower-elementary teachers. As there are non-trivial differences between the two groups, this is not ideal (see section 4) but the groups are similar in the sense that they through their educational and professional choices have expressed an interest in the education of young children. The majority of teachers in both groups also tend to be municipal employees or employed by private providers funded by the municipalities.

3.4. National test in sciences and social sciences in grade 9

National tests for students in grade 9 were introduced in the three science subjects (biology, chemistry, and physics) in 2008/09. In order to reduce the workload, each school is only the subject to one of the tests in a specific year. The tests are held in the middle of the spring semester (late March or early April). In both the sciences and social sciences, the exact subject being tested is announced to the school approximately one month prior to the test date.

As the control group to teachers in the sciences, we use teachers in never-tested practical subjects (music, arts, sports, home economics, technology, and crafts), and teachers in foreign languages, except for English. The motivation for not using teachers in core subjects (English, Swedish, and mathematics) is that tests in these subjects were introduced in 2003/04. Hence, there might be lingering consequences of these tests that affect the control group. As made clear in section 4, there are some distinct differences between the treatment and control groups in this setting.

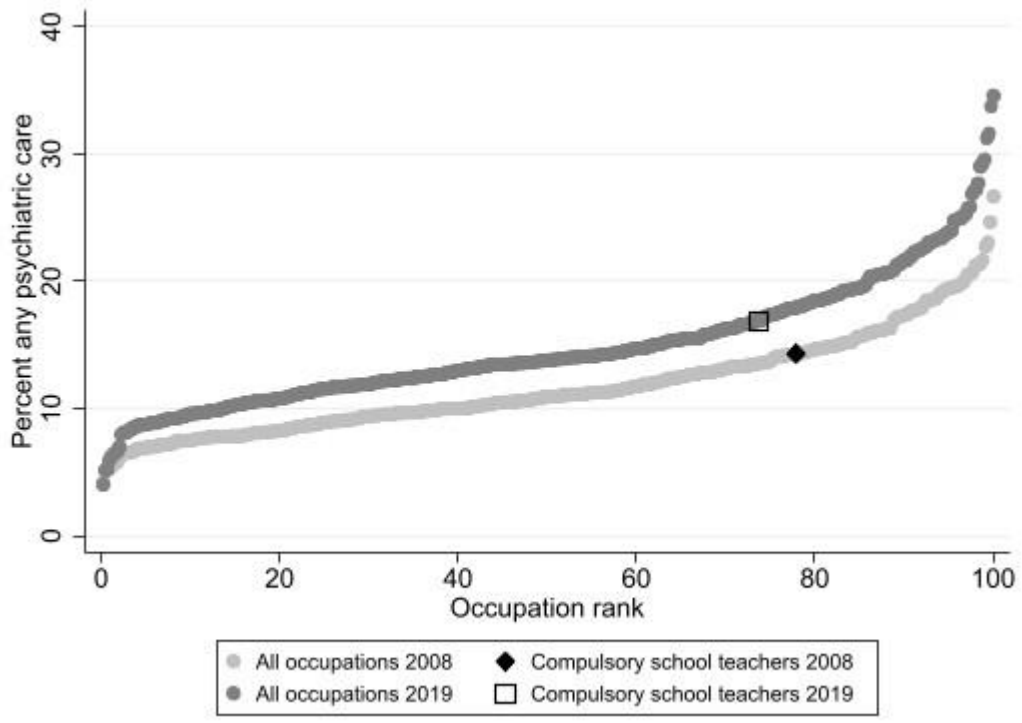
National tests for students in grade 9 were introduced in the four social science subjects (civics, geography, history, and religion) in 2012/13. As for the science subjects, each school is only administering one of the tests in a specific year and the tests are held in late April. As the control group to teachers in science subjects, we use teachers in always-tested core subjects (English, Swedish, and mathematics), never-tested practical subjects, and teachers in foreign languages. As for teachers in science subjects, there are non-trivial differences in the characteristics between the treatment and control groups.

4. Results

In order to put the health status of Swedish teachers in perspective, we start by examining the utilization of mental health care services and sick leave among compulsory school teachers compared to other occupations in 2008 and 2019. Figure 2 displays the share of workers who use some mental healthcare services by 4-digit occupational code, plotted against the occupational rank of utilization. The figure shows a general increase in the use of mental health care services, measured by either being in contact with health care for mental health problems or being prescribed a drug, throughout the distribution. The black diamond and square symbols represent the development of these outcomes among teachers, showing an increase in the use of mental health care services from 14.3 percent of the teachers in 2008 and 16.9 percent in 2019. While this is an increase in absolute numbers, mental healthcare use among teachers has decreased slightly relative to other occupations.

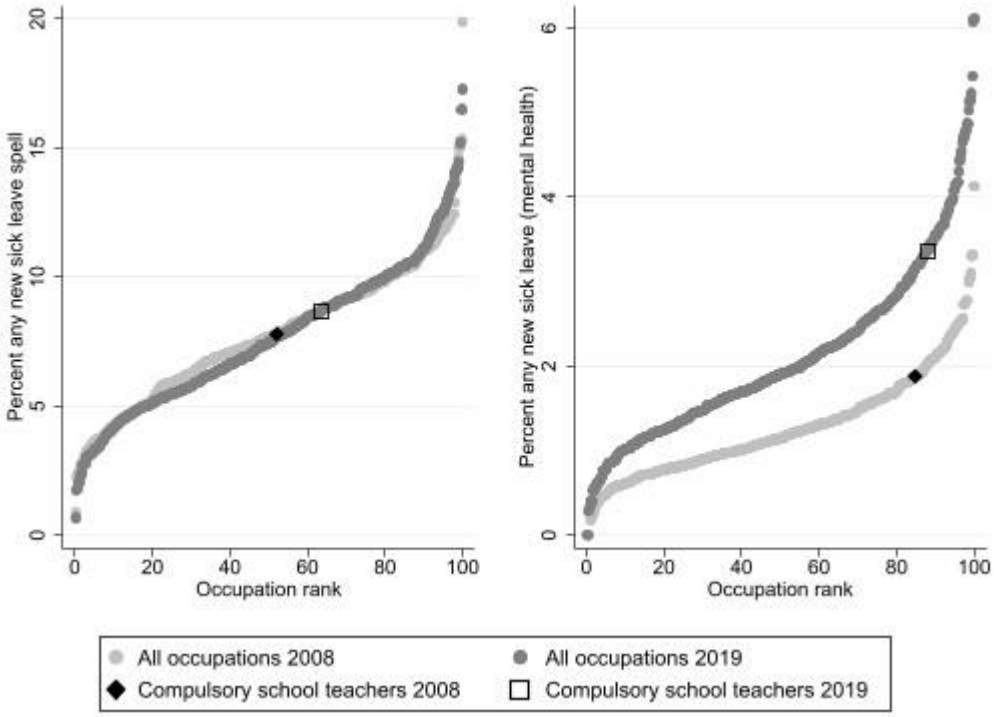
Figure 3 show the distribution of new sick leave spells longer than 2 weeks (left panel) and sick leave spells due to mental health problems (right panel). The level of total sick leaves is stable over time, but there has been a significant increase in the share of sick leave due to mental health problem. The increase among teacher is also slightly larger compared to other occupations. In 2008 1.9 percent of the teachers had at least one new sick leave spell for mental health problems, in 2019 the share was 3.4 percent. To conclude, mental health problems is an increasing problem in society, and this is also evident for teachers. In relation to other occupations, the results are mixed; teachers have fallen slightly in the rank of utilization of mental healthcare services and risen in the rank of sick leave.

Figure 2 Share of workforce using any mental health care services by occupation



Note: Occupations 4-digit SSYK96 codes with at least 100 individuals.

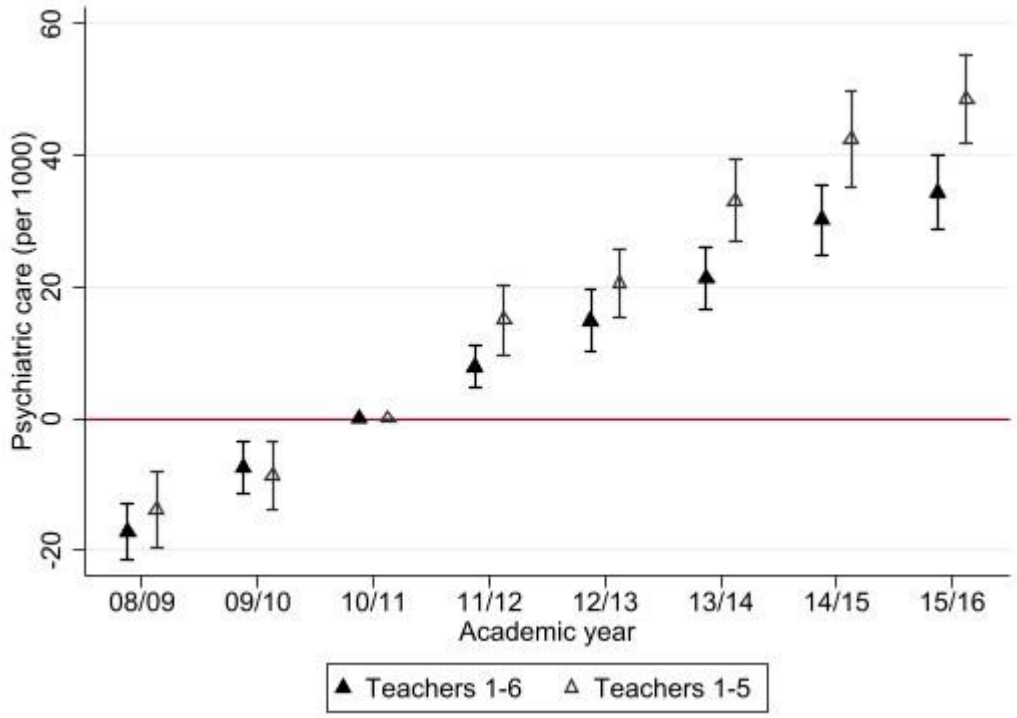
Figure 3 Share of workforce with a new sick leave spell due to any cause and due to mental health problems, by occupation



4.1. National tests in year 6

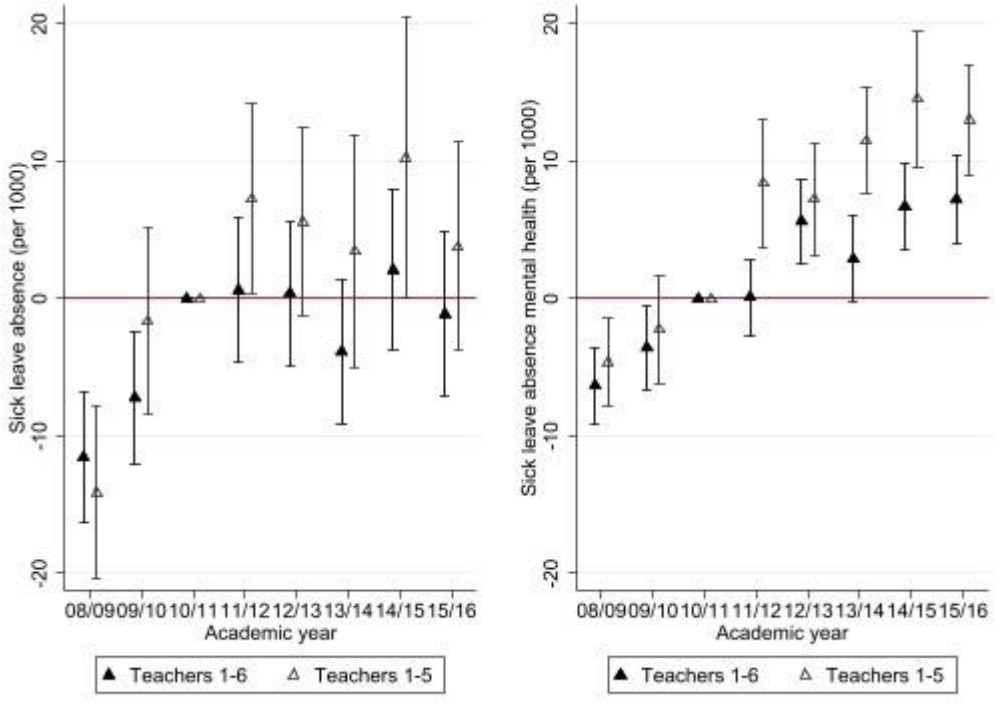
In this section, we present the findings of our analysis of the impact of national tests in grade 6 on teachers. Before turning to the results from estimation of the difference-in-differences model, it is useful to examine the trends in mental healthcare use and sick leave among teachers in 1-5 schools respectively 1-6 schools. Figure 4 illustrates that the use of mental healthcare services has increased significantly in both groups over time, although slightly greater increases observed among teachers in 1-5 schools. These results suggest that the implementation of standardized test had a negative impact on mental health. Figure 5 shows the trends in total sick leave (left panel) and sick leave due to mental health problems (right panel). The rates of sick leave increase in the first three years but then stabilize, whereas sick leave due to mental health problems increases over the entire period, similar to the trends in the use of mental health services. Similar to the trends in mental health care use, a visual inspection suggest that sick leave may have increased more among teachers in 1-5 schools.

Figure 4. Trends in mental healthcare utilization among upper-elementary teachers



Note: Use of mental healthcare services September to August each respective academic year, using 2010/11 as the reference year. The figure show estimates from separate linear regressions for upper-elementary teachers in schools with school years 1-5 respectively school years 1-6 on year fixed effects. Standard errors are clustered at municipality level. 95% CI indicated.

Figure 5. Trends in total sickness and due to mental health issues among upper-elementary teachers

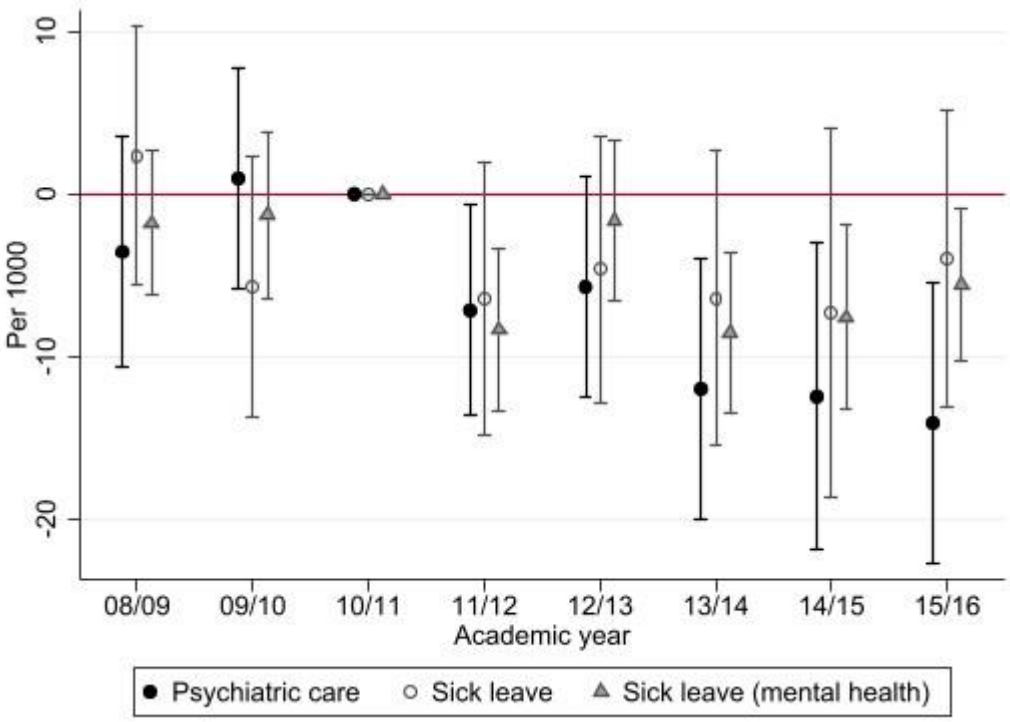


Note: Total sick leave (left panel) and sick leave due to mental health issues (left panel). Sick leave are measured for the period September to August each respective academic year, using 2010/11 as the reference year. The figure show estimates from separate linear regressions for upper-elementary teachers in schools with school years 1-5 respectively school years 1-6 on year fixed effects. Standard errors are clustered at municipality level. 95% CI indicated.

The difference-in-differences estimates are displayed in the figures below, with additional details provided in the tables found in the appendix. Figure 6 presents the results on mental health care utilization, total sick leave and sick leave due to mental health problems among teachers in schools which has to handle national tests versus those without. The results indicate a slight decrease in the use of mental health services among teachers in schools that conducted national test, with the difference between the two increasing over time. The year after the reform the relative decrease is 5.4 percent (0.7 percentage points) compared to pre-reform levels. Three years after the reform, the relative decrease is 9.3 percent (1.2 percentage points). Likewise, sick leave due to mental health problems seems to decrease. This effect large, about 40 percent (0.8 percentage points) compared to the pre-reform year. The estimated coefficients for total sick leave and sick leave due to mental health is very similar but the standard errors for total sick leave is larger, suggesting there is, as expected, more noise when including sick leave due to any cause. From here on we will therefore focus on the sick leave for mental health problems. The effect seems to become larger at the end of the period. This may partially be attributed to the introduction of grades in 6^{th} grade in 2013.

The validity of our findings hinges on the comparability of teachers in schools with grades 1-6 and those in schools with grades 1-5. Summary statistics provided in Table A1 of the appendix reveal some minor differences between these groups. For instance, a slightly higher percentage of teachers in 1-6 schools are men compared to those 1-5 grade schools (19 vs 17 percent). There are however no differences in our health outcomes. To account for potential compositional effects, the results displayed in the figures include controls for age category, sex and some information on the partner interacted with time fixed effect. Results without controls is available in appendix table A2. As can be seen in the table, including controls has essentially no effect on the estimates, suggesting these differences are unimportant. Moreover, the identifying assumption of parallel trends seems to be fulfilled since there is no evidence of diverging trends before 2011/12.

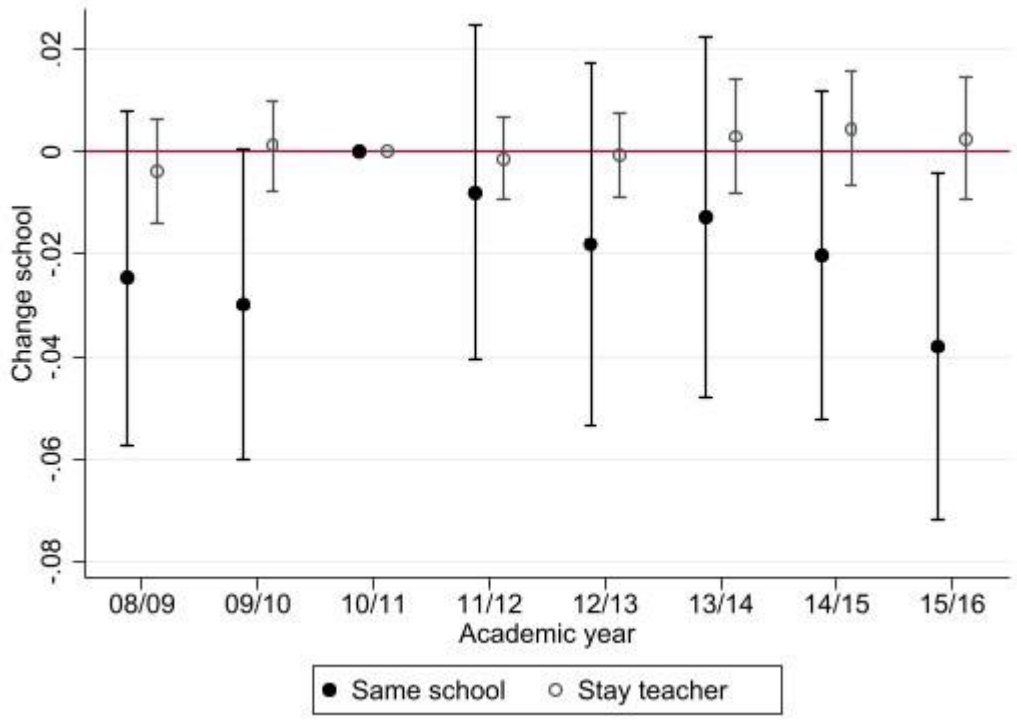
Figure 6. DID estimates of introducing national tests in grade 6 on the mental health care utilization and sickness absence



Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls (interacted with year fixed effects) are age category, sex, education category, if have partner and income percentile of partner. Standard errors are clustered at the municipality level. 95% CI indicated.

Next, we examine the effect of the introduction of national test on two measures of teacher turnover: i) whether a teacher remains in the same school the following year, and ii) whether a teacher continues working as a teacher the next year. The results presented in figure 7 show no impact of the introduction of national tests on teacher turnover.

Figure 7. DID estimates of national tests in grade 6 on changing workplace and working as teacher



Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

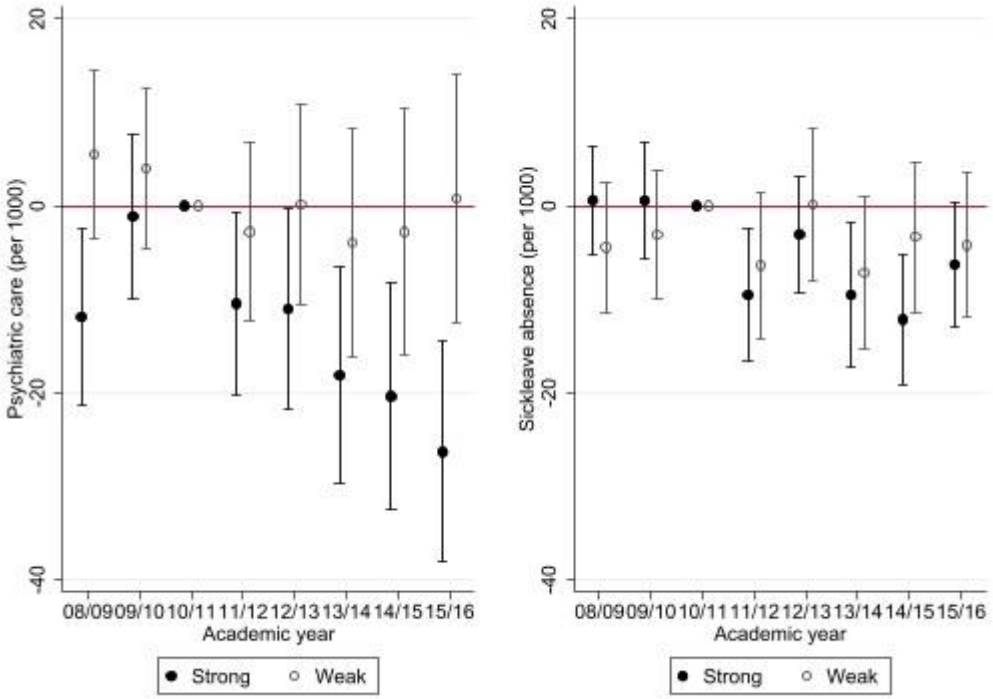
4.2. Heterogenous effects for grade 6 tests

As previously discussed, the impact of national tests on the work environment and stress may vary according to student composition, teacher age and whether schools are private or publicly operated. To examine this further, we estimate separate models for different groups, and present the results in Figures 8-10. Tables A4-A6 in the appendix contains estimates of the difference in the treatment effect between groups, specifically strong versus weak student background, older versus younger teachers and private versus public schools.

We find that the negative impact on the utilization of mental health services appears to be driven by schools with students who have higher predicted grades. Among schools below the median of predicted grades, healthcare use is stable compared to the control group. This pattern is in line with an increased focus on standardized outcomes being relatively beneficial to the work environment at higher performing schools. The increased workload and potential stress from having students fail are less common at such schools and having high test-scores also allows for higher subject-grades.[4] The difference in sick leave between the two types of schools is much less pronounced.

[4] Since subject-grades were introduced after the reform to the testing system, we cannot study if the reform affected the distribution of grades differently at schools with how and low expected achievement. Teacher-set grades are however known to be strongly influenced by the student composition and the same level of achievement tends to render students higher grades when peers are low-performing (e.g. Calsamiglia and Loviglio, 2009; for Swedish evidence, see Skolverket, 2019).

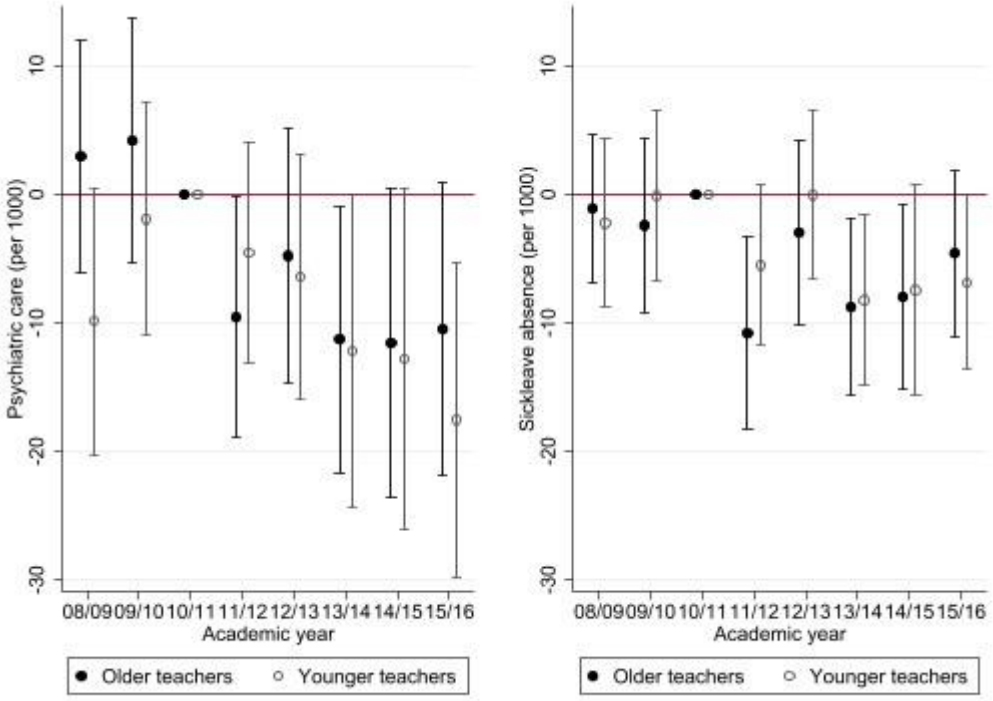
Figure 8. DID estimates on mental health and sickness absence due to mental health problems among teacher in schools with students with strong respectively weak family background



Note: Difference-in-differences estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

As discussed in relation to the theoretical work by Besley and Ghatak (2019), teachers are to some extent self-selected with respect to the incentives in place. The introduction of new types of evaluation might therefore impact older and younger teachers differently. We analyze this by splitting the sample along the median age in the year prior to the reform. As can be seen in Figure 9, however, there estimates show no substantive differences between the two groups. Admittedly, this is a relatively weak test of the hypothesis derived from Besley and Ghatak as all teachers in our sample would have based their occupational choice on a system without standardized testing.

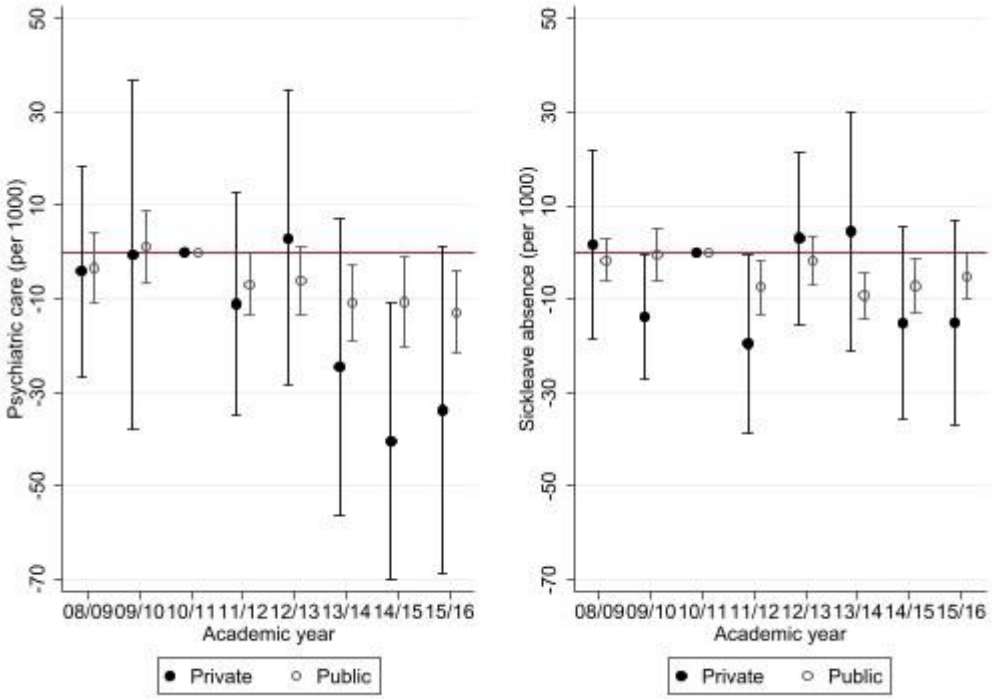
Figure 9. DID estimates on mental health and sickness absence due to mental health problems among older and younger teachers



Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

Sweden has a non-trivial fraction of private, voucher funded, school providers. In Figure 10 we test if the responses differ between teachers at privately and publicly run schools. While the theoretical prediction is less clear for this test, it is plausible that public and private employers respond differently when a new system of external evaluation is introduced. There is some indication that the decrease in mental health use is more pronounced among teachers at privately run schools, particularly towards the end of the period. However, given that only a small fraction of the teachers work in such schools, the standard errors are large and the differences in estimates are not statistically significant. For sick leave, the differences between teachers at private and public schools are smaller and, again, not statistically significant.

Figure 10. DID estimates on mental health and sick leave due to mental health problems among teacher in private and public schools

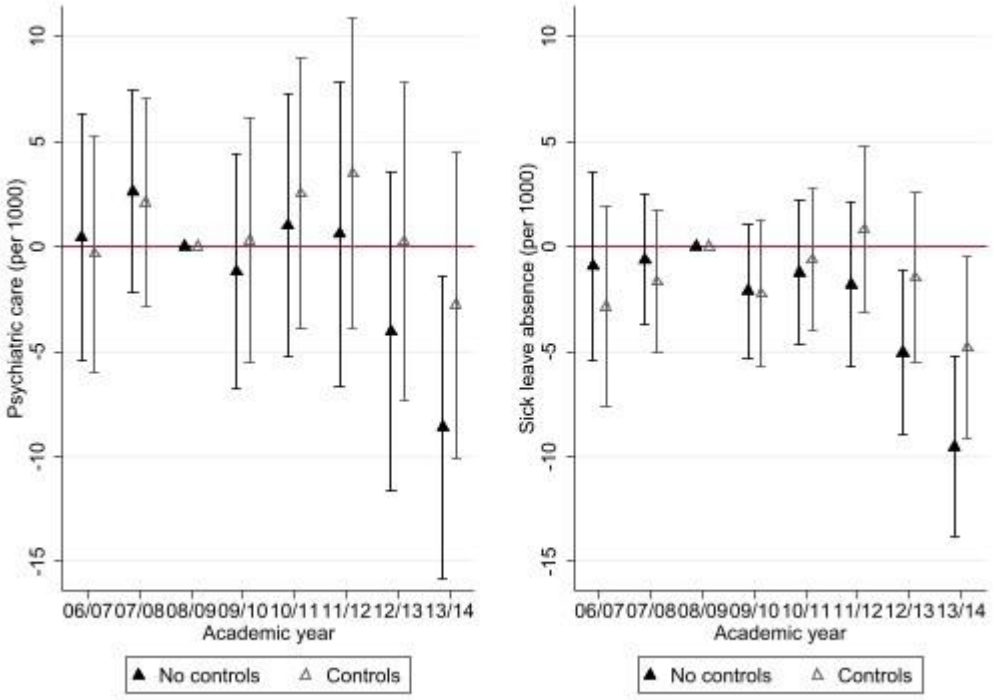


Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

4.3. National tests in grade 3

Our results indicate that the implementation of national tests in grade 6 did not have any detrimental effects on the mental health of teachers. If anything, the results suggest the opposite. In this section we show results for the implementation of national tests in grade 3. As previously discussed, there is no group of elementary school teachers that was not affected by the reform, so we compare them with preschool teachers. It is important to note that preschool and elementary school teacher differ in several ways, such as education level and the health outcomes, as shown in Table A1 in appendix. Figure 11 presents the difference-in-differences estimates from the model with and without control variables (interacted with time effects). In contrast to the results presented for grade 6 teachers, including control variables affect the estimates, suggesting that the compositional differences between the groups matter. With this caveat in mind, the results for the two outcomes are similar and indicate no significant effect of introducing national tests in grade 3 on the mental health of elementary school teachers.

Figure 11. DID estimates of introducing national tests in grade 3 mental health and sick leave

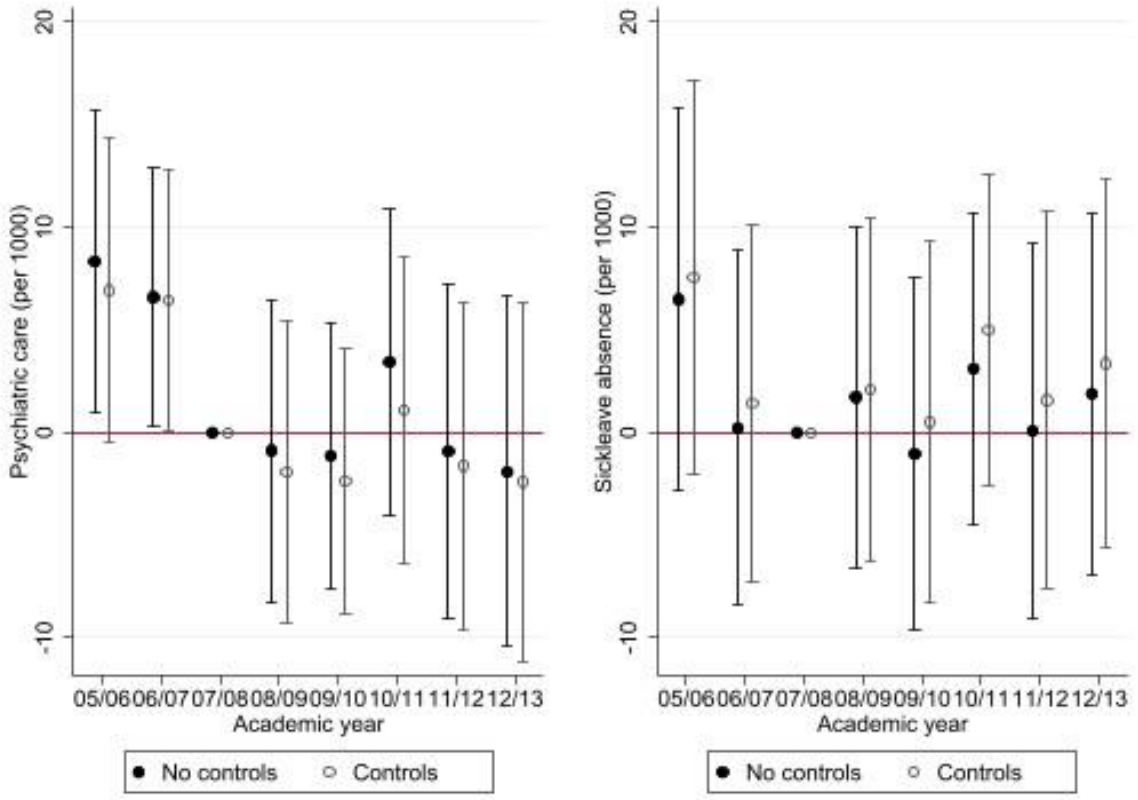


Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

4.4. National tests in Science subjects in year 9

National tests in science subjects were introduced in 2009. Figure 12 presents the difference-in-differences estimates for utilization of mental health services (left panel) and sick leave due to mental health problems (right panel). Our results indicate that, similar to the previous cases, the introduction of national tests in science subjects did not have any significant effects on mental health. However, it is worth noting that for utilization of mental health services the trend for teachers in science subjects appears to be decreasing in the years leading up to the reform, in compared to upper-secondary school teachers in other subjects. This may suggest that the assumption parallel trend is not fulfilled, and should be taken into consideration when interpreting the results.

Figure 12. DID estimates of introducing national tests in grade 9 sciences in grade 9 on mental health and sick leave due to mental health problems

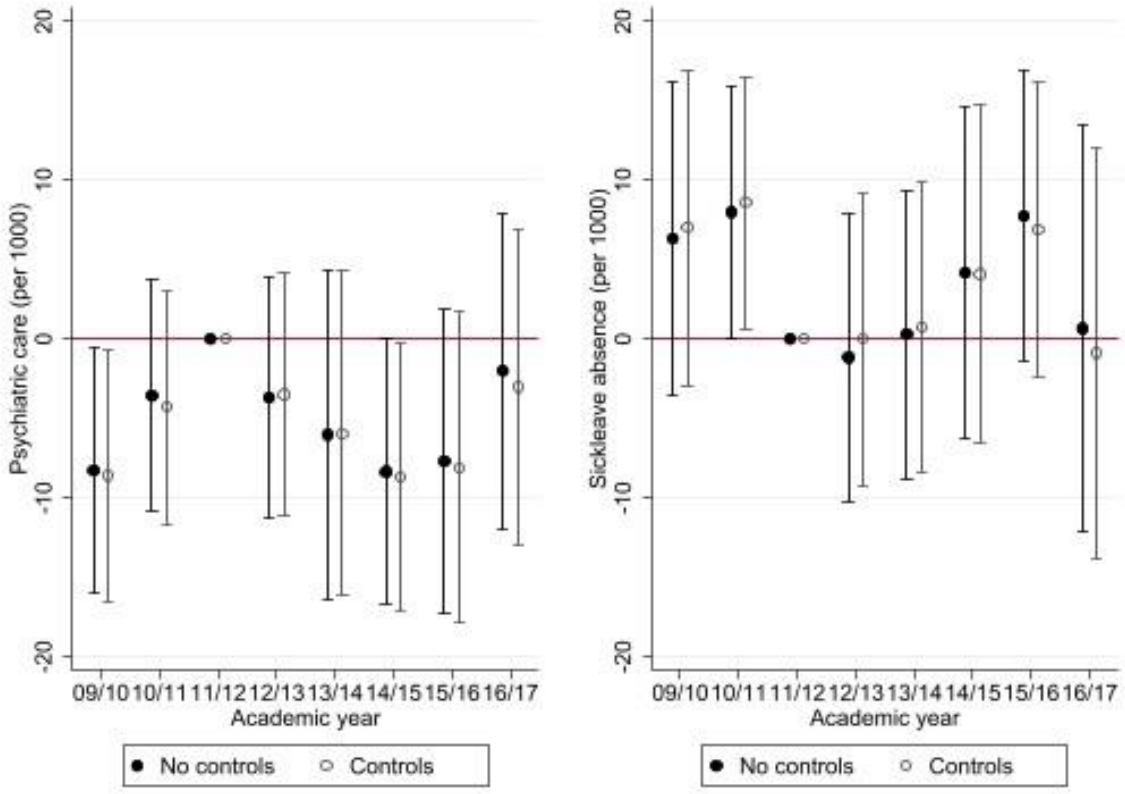


Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

4.5. National tests in the social science subjects in year 9

Our final experiment examines the implementation of national tests in social science subjects in 2013. The results, presented in figure 13, indicate that the introduction of national tests have no effect on teacher mental health. However, a concern that in this experiment, is that the assumption of parallel trend may not hold since there seem to be some differences in the trend before the reform.

Figure 13. DID estimates of introducing national tests in grade 9 in social sciences on mental health and sickness absence



Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls are age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. 95% CI indicated.

5. Discussion and conclusion

The increased use of standardized testing is widely claimed to be a contributing factor to teacher stress and burnout. While some correlational evidence support this view, quasi-experimental evidence to support such claims are the best of our knowledge, non-existent. We address the issue by studying the introduction and expansion of standardized testing in Sweden using difference-in-differences approaches and measures of healthcare use and sick leave. In none of the settings studied, we find any indication of testing to result in worsened teacher health or wellbeing. When tests were introduced in grade 6, measures of mental healthcare use based on diagnoses from hospitals and prescriptions of antidepressants and tranquilizers, fell by close to 10 percent (around 1 percentage point). For absence from work due to mental health issues, the estimates show a similar size effect in percentage points, which compared to the mean imply a 40 percent decrease. Thus, there is a substantial reduction in sick leave due to mental health problems. Further, estimates for teacher turnover suggest a null-effect. A more detailed analysis shows that the positive impact on health is concentrated to teachers working at schools with students from a relatively strong socioeconomic background. For other outcomes, no meaningful heterogeneities are identified. When studying reforms to the testing system for grades 3 and 9, we can to the very the least rule out substantive detrimental effects. However, the parallel trends assumption is more questionable in these settings.

The quasi-experimental literature on test-based accountability has so far primarily focused on the impact of test-based accountability on teacher mobility and sorting across schools. Our main contribution is adding to existing research by providing evidence on direct measures of mental health. As our results suggest a relative improvement in mental health, among teachers at schools with stronger students, our results are in line with existing findings. Our results question the commonly held view that testing contribute to a lower workplace quality and deteriorating teacher mental health. It should be stressed, however, that no centralized system of incentives is tied to the results on Swedish standardized tests and the extent to which local incentives are used in unknown.

References

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. Journal of teacher Education, 51(5), 384-397.

Bergbauer, A. B., Hanushek, E. A., & Woessmann, L. (2021). Testing. Forthcoming in Journal of Human Resources.

Berlak, A. and Berlak, H. (1981). Dilemmas of schooling. London: Methuen & Co. Ltd.

Besley, T., & Ghatak, M. (2018). Prosocial motivation and incentives. Annual Review of Economics, 10, 411-438.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2008). The impact of assessment and accountability on teacher recruitment and retention: Are there unintended consequences?. Public Finance Review, 36(1), 88-111.

Calsamiglia, C., & Loviglio, A. (2019). Grading on a curve: When having good peers is not good. Economics of Education Review, 73, 101916.

Clotfelter, C. T., Ladd, H. F., Vigdor, J. L., & Diaz, R. A. (2004). Do school accountability systems make it more difficult for low-performing schools to attract and retain high-quality teachers?. Journal of Policy Analysis and Management, 23(2), 251-271.

Dizon-Ross, R. (2020). How does school accountability affect teachers? Evidence from New York City. Journal of Human Resources, 55(1), 76-118.

EASHW (2022). Education – evidence from the European Survey of Enterprises on New and Emerging Risks. European Agency for Safety and Health at Work.

Feng, L., Figlio, D., & Sass, T. (2018). School accountability and teacher mobility. Journal of Urban Economics, 103, 1-17.

Figlio, D., & Loeb, S. (2011). School accountability. Handbook of the Economics of Education, 3, 383-421.

Frostenson, M. (2015). Organisatoriska åtgärder på skolnivå till följd av lärarlegitimationsreformen. Institutet för arbetsmarknads-och utbildningspolitisk utvärdering Rapport 2015:3.

Gjefsen, H. M., & Gunnes, T. (2020). How school accountability affects teacher mobility and teacher sorting. Education Economics, 28(5), 455-473.

Green, C. P., Nyhus, O., & Salvanes, K. (2022). How does testing young children influence educational attainment and well-being? IZA DP 15686

Heissel, J. A., Adam, E. K., Doleac, J. L., Figlio, D. N., & Meer, J. (2021). Testing, stress, and performance: how students respond physiologically to high-stakes testing. Education Finance and Policy, 16(2), 183-208.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. The Journal of Law, Economics, and Organization, 7, 24-52.

Högberg, B., Lindgren, J., Johansson, K., Strandh, M., & Petersen, S. (2021). Consequences of school grading systems on adolescent health: evidence from a Swedish school reform. Journal of Education Policy, 36(1), 84-106.

Iriate Redín, C., & Erro-Garcés, A. (2020). Stress in teaching professionals across Europe. International Journal of Educational Research, 103, 101623.

Jerrim, J. and Sims, S. (2022), "School accountability and teacher stress: international evidence from the OECD TALIS study", Educational Assessment, Evaluation and Accountability, 34:5-32

Johnsson, JV. and EM. Hall (1988), "Job strain, work place social support, and cardiovascular disease: a cross-sectional study of a random sample of the Swedish working population", American Journal of Public Health, 78(10): 1336-1342.

Karasek, R. (1979), "Job demands, job decision latitude and mental strain: Implications for job redesign", Administrative Science Quarterly, 24:2, 285-308.

Luna, C., & Turner, C. L. (2001). The impact of the MCAS: Teachers talk about high-stakes testing. The English Journal, 91(1), 79-87.

Miller, R. T., Murnane, R. J., & Willett, J. B. (2008). Do teacher absences impact student achievement? Longitudinal evidence from one urban school district. Educational Evaluation and Policy Analysis, 30(2), 181-200.

OECD (2015). Fit Mind, Fit job: from evidence to practice in mental health and work. In Fit Mind, Fit Job: From Evidence to Practice in Mental Health and Work. OECD/ODCE.

OECD (2021). Fitter Minds, Fitter Jobs: Mental Health and Work. From Awareness to Change in Integrated Mental Health, Skills and Work Policies.

Reback, R., Rockoff, J., & Schwartz, H. L. (2014). Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind. American Economic Journal: Economic Policy, 6(3), 207-241.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. American educational research journal, 50(1), 4-36.

Siegrist, J., Siegrist, K. and I. Weber (1986), "Sociological concepts in the etiology of chronic disease: The case of ischemic heart disease", Social Science & Medicine, 22:247, 104–126.

Sievertsen, H. H. (2022). Assessments in Education. arXiv preprint arXiv:2208.05826.

Ulrich, C., O'Donell, P., Farrar, A., Danis, M. and C. Grady (2007), "Ethical climate, etics stress, and the job satisfaction of nurses and social workers in the United States", Social Science and Medicine 65(8): 1708-719.

Skolverket (2016). Nationella proven i grundskolans årskurs 6 och 9. Rapport 447.

Skolverket (2019). Analyser av likvärdig betygssättning mellan elevgrupper och skolor. Rapport 475.

Vlachos, J. (2019), "Trust-based evaluation in a market-oriented school system". In Dahlstedt, M., & Fejes, A. (Eds.), Neoliberalism and Market Forces in Education: Lessons from Sweden. Routledge.

Von der Embse, N. P., Schoemann, A. M., Kilgus, S. P., Wicoff, M., & Bowler, M. (2017). The influence of test-based accountability policies on teacher stress and instructional practices: A moderated mediation model. Educational Psychology, 37(3), 312-331.

von der Embse, N., Ryan, S. V., Gibbs, T., & Mankin, A. (2019). Teacher stress interventions: A systematic review. Psychology in the Schools, 56(8), 1328-1343.

Appendix

Table A1 Summary statistics teachers at 1-5 vs 1-6 schools

Teachers in 1-5 schools vs 1-6 schools				
	1-5 schools	1-6 schools	Difference	P-value
Age	46.70	46.98	-0.27	0.038*
Man	0.17	0.19	-0.02	0.000***
Number of children age 0-19	1.03	1.06	-0.03	0.052
Higher education >2 years	0.81	0.83	-0.03	0.000***
Have a partner	0.73	0.75	-0.02	0.001***
Partner income percentile, by age, sex and year	62.91	61.08	1.83	0.000***
Psychiatric care	135.38	129.02	6.36	0.123
New sick leave spell >14 days	89.03	88.48	0.55	0.874
New sick leave spell with mental health diagnoses >14 days	20.84	24.27	-3.43	0.062
Teacher next year	0.90	0.90	0.00	0.675
Work at same school net year	0.67	0.71	-0.04	0.000***
Observations	10075	19857		

Elementary school teachers vs preschool teachers				
	Teachers	Preschool	Difference	P-value
Age	51.16	45.03	6.13	0.000***
Man	0.04	0.07	-0.04	0.000***
Number of children age 0-19	0.77	1.03	-0.25	0.000***
Higher education >2 years	0.94	0.39	0.55	0.000***
Have a partner	0.76	0.69	0.07	0.000***
Partner income percentile, by age, sex and year	61.82	59.31	2.51	0.000***
Psychiatric care	141.89	133.45	8.45	0.012*
New sick leave spell >14 days	83.33	117.33	-33.99	0.000***
New sick leave spell with mental health diagnoses >14 days	19.13	24.41	-5.28	0.000***
Observations	11868	83119		

Teachers in lower secondary schools non-science vs science subjects				
	(1)			
	Non-science	Science	Difference	P-value
Age	45.24	44.49	0.75	0.000***
Man	0.31	0.46	-0.15	0.000***
Number of children age 0-19	0.98	1.11	-0.13	0.000***
Higher education >2 years	0.79	0.93	-0.14	0.000***
Have a partner	0.66	0.71	-0.05	0.000***
Partner income percentile, by age, sex and year	59.97	61.95	-1.98	0.000***
Psychiatric care	119.19	84.20	35.00	0.000***
New sick leave spell >14 days	78.71	63.98	14.73	0.000***
New sick leave spell with mental health diagnoses >14 days	17.54	15.79	1.75	0.319
Observations	21662	7221		

Teachers non-social science vs social science subjects	Non-science	Science	Difference	P-value
Age	46.04	44.63	1.41	0.000***
Man	0.33	0.44	-0.11	0.000***
Number of children age 0-19	1.03	1.05	-0.02	0.181
Higher education >2 years	0.86	0.96	-0.10	0.000***
Have a partner	0.68	0.68	0.00	0.764
Partner income percentile, by age, sex and year	60.97	62.93	-1.96	0.000***
Psychiatric care	122.58	125.16	-2.58	0.603
New sick leave spell >14 days	68.49	53.70	14.79	0.000***
New sick leave spell with mental health diagnoses >14 days	20.95	17.59	3.37	0.115
Observations	23960	5345		

Table A2 The impact of the reform on various outcomes

	(1) Mental health	(2) Mental health	(3) Sick leave (total)	(4) Sick leave (total)	(5) Sick leave (mental)	(6) Sick leave (mental)
	b/se	b/se	b/se	b/se	b/se	b/se
Year 2009	-3.300	-3.574	2.597	2.330	-1.692	-1.835
	(3.580)	(3.577)	(3.997)	(4.031)	(2.257)	(2.262)
Year 2010	1.184	0.941	-5.615	-5.720	-1.290	-1.335
	(3.418)	(3.429)	(4.082)	(4.101)	(2.607)	(2.601)
Year 2011	Ref	Ref	Ref	Ref	Ref	Ref
Year 2012	-7.105**	-7.166**	-6.687	-6.483	-8.253***	-8.382***
	(3.333)	(3.312)	(4.261)	(4.251)	(2.567)	(2.543)
Year 2013	-5.661	-5.756*	-5.183	-4.667	-1.589	-1.678
	(3.448)	(3.442)	(4.267)	(4.164)	(2.471)	(2.495)
Year 2014	-11.836***	-12.007***	-7.290	-6.418	-8.568***	-8.575***
	(4.077)	(4.093)	(4.991)	(4.627)	(2.501)	(2.517)
Year 2015	-12.313**	-12.466**	-8.173	-7.358	-7.807***	-7.632***
	(4.817)	(4.822)	(6.354)	(5.784)	(2.813)	(2.882)
Year 2016	-14.088***	-14.108***	-4.913	-4.025	-5.784**	-5.600**
	(4.335)	(4.388)	(4.991)	(4.633)	(2.390)	(2.383)
Controls		Yes		Yes		yes
Mean dep. var 1-5	135.38	135.38	89.03	89.03		20.84
R-squared	0.003	0.023	0.000	0.008	0.001	0.005
Observations	238,662	238,662	238,662	238,662	238,662	238,662

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A3 The impact of the reform on teacher turnover

	(1) Same school b/se	(2) Same school b/se	(3) Work as teacher b/se	(4) Work as teacher b/se
Year 2009	-0.019	-0.025	-0.000	-0.004
	(0.017)	(0.017)	(0.006)	(0.005)
Year 2010	-0.026*	-0.030*	0.002	0.001
	(0.015)	(0.015)	(0.005)	(0.004)
Year 2011	Ref	Ref	Ref	Ref
Year 2012	-0.008	-0.008	-0.002	-0.001
	(0.016)	(0.017)	(0.004)	(0.004)
Year 2013	-0.019	-0.018	-0.002	-0.001
	(0.018)	(0.018)	(0.005)	(0.004)
Year 2014	-0.014	-0.013	0.001	0.003
	(0.018)	(0.018)	(0.007)	(0.006)
Year 2015	-0.021	-0.020	0.002	0.005
	(0.016)	(0.016)	(0.007)	(0.006)
Year 2016	-0.037**	-0.038**	0.003	0.003
	(0.017)	(0.017)	(0.006)	(0.006)
Controls		Yes		Yes
Mean dep. var 1-5	0.67	0.67	0.90	0.90
R-squared	0.198	0.258	0.046	0.210
Observations	238,662	238,662	238,662	238,662

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A4: Schools with strong vs weak student background

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2009×Strong	-17.370***	-17.469***	5.091	4.962
	(6.075)	(6.070)	(4.585)	(4.569)
Year 2010×Strong	-5.063	-5.158	3.726	3.676
	(5.529)	(5.527)	(4.084)	(4.062)
Year 2011×Strong	Ref	Ref	Ref	Ref
Year 2012×Strong	-7.671	-7.754	-3.144	-3.459
	(7.157)	(7.161)	(5.525)	(5.488)
Year 2013×Strong	-11.179	-11.213	-3.165	-2.956
	(8.446)	(8.391)	(5.314)	(5.282)
Year 2014×Strong	-14.258	-14.838	-2.343	-2.174
	(9.124)	(9.227)	(6.280)	(6.181)

Year 2015×Strong	-17.592**	-18.265**	-8.797*	-8.675*
	(7.994)	(7.924)	(5.038)	(4.991)
Year 2016×Strong	-27.104***	-27.496***	-2.048	-1.556
	(8.999)	(8.975)	(5.375)	(5.331)
Controls		Yes		Yes
Mean dep. var weak	137.11	137.11	22.02	22.02
Mean dep. var strong	134.14	134.14	19.99	19.99
R-squared	0.003	0.023	0.001	0.005
Observations	238,662	238,662	238,662	238,662

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A5: Older vs younger teachers

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2009×Older	12.861*	13.327*	1.167	1.241
	(6.760)	(6.787)	(4.396)	(4.398)
Year 2010× Older	6.136	6.542	-2.331	-2.120
	(6.483)	(6.501)	(4.434)	(4.434)
Year 2011× Older	Ref	Ref	Ref	Ref
Year 2012× Older	-5.052	-5.012	-5.292	-5.474
	(6.313)	(6.350)	(4.807)	(4.834)
Year 2013× Older	1.650	1.947	-2.908	-2.846
	(7.050)	(6.997)	(4.978)	(4.986)
Year 2014× Older	0.944	1.193	-0.567	-0.573
	(8.108)	(8.328)	(4.766)	(4.792)
Year 2015× Older	1.288	1.478	-0.532	-0.856
	(8.497)	(8.723)	(5.377)	(5.406)
Year 2016× Older	7.052	7.155	2.280	2.301
	(8.252)	(8.506)	(4.787)	(4.867)
		Yes		Yes
Mean dep. var young	93.85	93.85	18.41	18.41
Mean dep. var old	175.40	175.40	23.19	23.19
R-squared	0.012	0.023	0.002	0.005
Observations	238,662	238,662	238,662	238,662

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A5: Private vs public schools

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2009×Private	-0.683	-7.128	3.410	4.350

	(12.199)	(12.124)	(10.446)	(9.960)
Year 2010× Private	-1.819	-6.282	-13.185*	-13.264*
	(20.482)	(19.798)	(7.379)	(7.481)
Year 2011× Private	Ref	Ref	Ref	Ref
Year 2012× Private	-4.246	-7.586	-12.074	-11.481
	(12.503)	(11.965)	(10.736)	(10.319)
Year 2013× Private	9.149	6.681	4.892	6.195
	(16.877)	(16.068)	(10.008)	(10.033)
Year 2014× Private	-13.637	-18.673	13.710	14.578
	(16.520)	(16.068)	(13.329)	(12.509)
Year 2015× Private	-29.868**	-34.501**	-7.973	-5.557
	(15.119)	(14.869)	(10.382)	(10.623)
Year 2016× Private	-20.991	-23.269	-9.894	-8.863
	(17.989)	(16.429)	(11.763)	(11.884)
Controls		Yes		Yes
Mean dep. var public	136.85	136.85	21.41	21.41
Mean dep. var private	115.33	115.33	13.14	13.14
R-squared	0.003	0.023	0.001	0.005
Observations	238,662	238,662	238,662	238,662

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A7: Impact of national tests in grade 3

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2007	0.430	-0.373	-0.929	-2.869
	(2.978)	(2.870)	(2.266)	(2.414)
Year 2008	2.600	2.100	-0.640	-1.661
	(2.448)	(2.531)	(1.581)	(1.710)
Year 2009	Ref	Ref	0.000	0.000
			(.)	(.)
Year 2010	-1.192	0.288	-2.111	-2.224
	(2.839)	(2.937)	(1.629)	(1.760)
Year 2011	1.002	2.520	-1.248	-0.636
	(3.186)	(3.259)	(1.734)	(1.721)
Year 2012	0.608	3.455	-1.811	0.812
	(3.683)	(3.753)	(1.995)	(2.002)
Year 2013	-4.057	0.229	-5.048**	-1.489
	(3.852)	(3.837)	(2.001)	(2.056)
Year 2014	-8.606**	-2.768	-9.544***	-4.837**
	(3.668)	(3.706)	(2.176)	(2.205)
		Yes		Yes
Mean dep. var pre-school teachers	133.45	133.45	24.41	24.41
R-squared	0.002	0.021	0.001	0.004

Observations	757,294	757,294	757,294	757,294

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

Table A8: Impact of national tests in science subjects in grade 9

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2006	8.329**	6.919*	-3.752*	-3.254
	(3.716)	(3.762)	(2.160)	(2.200)
Year 2007	6.596**	6.445**	-1.501	-0.572
	(3.226)	(3.249)	(2.028)	(2.120)
Year 2008	Ref	Ref	Ref	Ref
Year 2009	-0.885	-1.961	-0.611	-0.520
	(3.742)	(3.731)	(2.177)	(2.206)
Year 2010	-1.134	-2.390	-5.154**	-4.798**
	(3.290)	(3.279)	(2.279)	(2.308)
Year 2011	3.446	1.088	-0.026	0.936
	(3.786)	(3.797)	(2.022)	(2.005)
Year 2012	-0.903	-1.639	-1.333	-1.221
	(4.120)	(4.067)	(2.471)	(2.482)
Year 2013	-1.924	-2.413	-0.149	0.542
	(4.333)	(4.461)	(2.534)	(2.609)
Controls		Yes		Yes
Mean dep. var	121.12	121.12	19.11	19.11
R-squared	0.004	0.030	0.000	0.004
Observations	230,232	230,232	230,232	230,232

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.

.

Table A9: Impact of national tests in social science subjects in grade 9

	(1) Mental health b/se	(2) Mental health b/se	(3) Sick leave b/se	(4) Sick leave b/se
Year 2010	-8.276**	-8.607**	1.323	1.722
	(3.930)	(4.027)	(2.368)	(2.439)
Year 2011	-3.578	-4.283	3.525	3.778
	(3.695)	(3.746)	(2.323)	(2.382)
Year 2012	Ref	Ref	Ref	Ref
Year 2013	-3.686	-3.487	2.994	3.395
	(3.878)	(3.900)	(2.938)	(2.957)
Year 2014	-6.024	-5.935	-0.083	0.341

	(5.276)	(5.212)	(3.007)	(3.026)
Year 2015	-8.346*	-8.700**	-1.722	-1.763
	(4.279)	(4.288)	(2.640)	(2.652)
Year 2016	-7.698	-8.079	-0.601	-0.876
	(4.851)	(4.985)	(2.602)	(2.653)
Year 2017	-2.010	-3.018	-0.675	-1.161
	(5.047)	(5.055)	(3.441)	(3.462)
		Yes		Yes
Mean dep. var	122.58	122.58	20.95	20.95
R-squared	0.003	0.022	0.001	0.005
Observations	233,609	233,609	233,609	233,609

Note: Difference-in-difference estimates from the linear regression model (1) specified in the text. Controls include: age category, sex, education category, if have partner and income percentile of partner. Standard errors cluster at municipality level. Standard errors in parentheses clustered at municipality level. *** significant at 1%, ** at 5%,* at 10%.