I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15686

How Does Testing Young Children Influence Educational Attainment and Well-Being?

Colin P. Green
Ole Henning Nyhus
Kari Vea Salvanes

OCTOBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15686

How Does Testing Young Children Influence Educational Attainment and Well-Being?

Colin P. Green
Norwegian University of Science and Technology and IZA

Ole Henning Nyhus
Norwegian University of Science and Technology Social Research

Kari Vea Salvanes
Nordic Institute for Studies in Innovation, Research and Education

OCTOBER 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

How Does Testing Young Children Influence Educational Attainment and Well-Being?*

How much young children should be tested and graded is a highly contentious issue in education policy. Opponents consider it detrimental to child mental health, leading to misaligned incentives in educational policy and having little if any redeeming impact on educational performance. Others see early testing of children as a necessary instrument for identifying early underachievement and educational targeting while incentivising schools to improve the educational performance of children. In practice, there is large crosscountry variation in testing regimes. We exploit random variation in test-taking in mathematics among early primary school children in Norway, a low testing environment. We examine two forms of testing, complex but low-stakes mathematics tests and relatively easy screening tests aimed at identifying children in need of educational assistance. In general, we demonstrate zero effects of testing exposure on later test score performance but benefits for screening tests on low-performing students. While we demonstrate no negative effects on student welfare, we do find an indication that testing improves aspects of teaching practices and students' perceptions of teacher feedback and engagement.

JEL Classification: I28, I24

Keywords: student assessment, testing, student achievement

Corresponding author:
Colin Green
Department of Economics
Norwegian University of Science and Technology
Klæbuveien 72
Trondheim
Norway

E-mail: colin.green@ntnu.no

* We thank project collaborators Simon Calmar Andersen, Hans Bonesrønning, Ester Bøckmann, Henning Finseraas, Ines Hardoy, Jon Marius Vaag Iversen, Vibeke Opheim, Astrid Marie Jorde Sandsør, and Pål Schøne, and participants at seminars and conferences (Scottish Economic Society 2022; LEER workshop 2021 on 'Experimental Evidence in Education Economics; Department of Economics at Norwegian University of Science and Technology; Educational Resources and Student Performance Workshop, Oslo 2022; Forskermøtet, Stavanger 2022) for comments on earlier drafts. This research is part of the 1+1 Project, supported by the Norwegian Research Council under Grant 256217.

1. Introduction

How much should young children be tested and graded is one of the most contentious issues in education policy. There are large international variations in practice, yet simple cross-country comparisons provide little guidance on likely differences in effects on educational attainment. One can, for instance, contrast the testing regimes of the countries that typically head the PISA and TIMMS tables, South Korea, Japan, and Finland. Countries that can be viewed as being at the extremes of testing practices. Opponents to testing young children, for example in the Nordic countries, view it as detrimental to children's mental health. They also raise concerns that it leads to misaligned incentives in educational policy and has little, if any, redeeming impact on educational performance. While in other countries, such as the UK and US, proponents see early testing of children as a necessary instrument for identifying early underachievement, educational targeting, and incentivising schools to improve the educational performance of children.

Despite the importance of this topic and the level of disagreement, there is relatively little evidence on the impact of testing on students and student outcomes. This reflects a range of factors, including a lack of within jurisdiction variation in testing practices and that the absence of testing often mechanically leads to a lack of readily available measures of educational attainment. Such issues make it especially difficult to examine the effects of testing in traditionally low-testing environments. What we know comes primarily from three sources and focuses primarily on older school children. In cross-country comparisons of testing regimes and their effects, Bergbauer et al. (2021) use PISA to demonstrate positive effects of testing are present only in lower-performing countries, while Högberg and Horn (2022) find adverse effects of high-stakes testing on stress analysing three waves of the European Health Behaviour in School-aged Children study. The second source of evidence comes from studies that looks at changes in specific testing regimes and school accountability within a given country, for instance No Child Left Behind (NCLB) in the US (Dee and Jacobs, 2011) and the removal of league tables in Wales (Burgess et al., 2013). A recent and small literature exploits events where testing did not occur in a specific year due to shocks to the testing environment in the UK (Murphy, Regan-Stansfield and Wyness, 2022) and Denmark (Anderson and Nielsen, 2020).

We return to this focusing on Norway, a country that, in common with other Nordic countries, is historically a no to low testing environment. The assignment of grades and examination of children has historically been avoided in Norway. As we discuss later, the first national primary school tests

in Math, Norwegian and English were only first introduced in grade 5 (age 10) in 2004.[1] There remains no general examination of educational attainment below this age. Our first aim is to examine how introducing testing in this environment influences primary school children's educational attainment.

We do so in two ways. First, we exploit a large multi-year RCT in Norway, the 1+1 project. The focus of this project was the introduction of additional small group math instruction for young primary school students (Bonesrønning et al., 2022). While this RCT was not designed to examine the effect of testing, a side product was the introduction of a standardised math tests to Norwegian children aged 7-9. As this was intended to provide detailed information on mathematical ability, a test that was both difficult and challenging was designed and used. As we discuss later, few students complete the entire test and there was large variation in student performance. These tests were implemented in all of the 159 schools of the RCT, both the 78 treatment schools and the 81 control schools. Importantly, the control schools received these tests but without additional small group instruction or any other treatment. As a result we use students in these schools to examine how testing influences later math attainment.

While these tests are the first general standardised tests to be taken by such young children in Norway, in 2008 a mandatory screening test in literacy and mathematics was introduced for all students in grade 3 (age 8). Unlike the 1+1 tests, which aim to provide an overall measure of a given child's mathematics attainment, these tests are screening tests aimed purely at identifying students with low educational attainment. As such they are relatively easy tests which provide little information other than whether students achieve some minimum standard. We exploit the introduction of this test to further examine effects of early testing on later attainment. Doing so provides complementary evidence on the effects of these forms of screening tests on student attainment and fits with one of the stated aims of early testing, identifying (and helping) students with educational needs.

A key contention is that testing young school children may generate a range of negative effects, where socio-psycho effects on children are a key concern. Even if there are academic benefits from early testing, as suggested by e.g. Roediger et al. (2011), these may be outweighed by negative effects. At the same time, there is a view that testing generates misaligned incentives for teachers and schools. We use novel data from national school environment surveys to explore these points. This data contains student reports on well-being, but also their perceptions of their interactions

[1] The initial tests were substantially revised in 2006 and national tests in their current form were introduced in 2007.

with teachers, teaching behaviour and feedback. We link this with our testing exposure from the 1+1 project to provide the first evidence on how testing affects these outcomes.

Together, this allows us to provide a range of evidence on the effects of the introduction of testing from a zero-testing baseline, evidence that we argue is missing from the current literature. Most directly, does early mathematics testing affect math attainment? Our main result is that there is a zero effect of the introduction of testing on the math attainment of students. This is robust to alternative identification strategies and does not appear to vary meaningfully by differences in test dosage. There is evidence that early screening tests have a positive effect of testing on later math attainment, concentrated in groups likely to face educational difficulties. We additionally demonstrate no negative effects on student welfare but present evidence that testing improves aspects of teaching practices and students' perceptions of teacher feedback and engagement. This suggests that, even in the absence of systematic positive effects on attainment, standardised testing may have a role in improving teaching and classroom practices. This runs counter to concerns that such testing introduces misaligned incentives unlikely to benefit students.

Our results suggest that introducing (difficult) low-stakes tests in low-testing environments is unlikely to meaningfully affect average educational performance. However, exams, particularly screening exams, may improve outcomes for students facing educational difficulties. There is little indication that testing has any negative consequences for student well-being, and may in fact lead to modest improvements in teacher-student interactions. Together, this provides evidence of benefits from introducing testing in these settings.

2. Institutional Background and Testing

2.1. The Norwegian school system

Norwegian children start school in August of the calendar year they turn 6. Compulsory education covers ten years of schooling divided into primary school (grades 1-7) and lower secondary school (grades 8-10). Compulsory schooling is free of charge, and nearly all students attend their local publicly owned school, with only approximately 4% of students attending private schools. All schools follow a national curriculum, share a common legal framework, and teachers must be accredited to teach in Norway.

A key feature is the lack of an emphasis on formal grades in primary school. For instance, entry into lower secondary schools is strictly a function of residential location. Prior to 2004, there was no formal grading, tests or exams in Norwegian primary schools, or any standardised testing before

the end of compulsory schooling. Pressure to introduce testing followed from Norway's poor performance in its first PISA participation in 2000.[2] In 2004, mandatory standardised exams in numeracy, reading and English as a foreign language were introduced for the fall of 5th grade. These tests are digital with numeracy and reading tests of 90 minutes, and English as a foreign language test of 60 minutes. The aim of these tests is to provide schools with information on student attainment levels in core skills and information used to improve teaching quality. Schools are instructed that exemptions from sitting the tests can only be granted for students in special needs education or for students learning Norwegian as a second language, and only if the schools consider that testing these children will not provide valuable information to teachers and schools. Between 5 and 6 percent of 5th graders are exempt from sitting these tests each year, while a further 1-2 percent are coded as "did not participate" as they were absent on the day of the test. Although students receive a formal grade on this test, it does not involve any stakes and is not used for decisions on, for example, grade retention. In fact, grade retention does not exist in Norway's compulsory education system.

2.2. The 1+1 Project and the introduction of mathematic exams

Our initial focus is on test-taking introduced through a large-scale multi-year RCT known as the "1+1 Project". This project was designed to examine the effect of additional math teaching resources (specifically an additional math teacher in pull-out small group instruction) targeting young primary school students.[3] To provide baseline information along with initial data on the impact of the intervention, students aged 7-9 in both treatment and control schools were made to sit a number of standardised (low stakes) digital tests in mathematics.[4]

Appendix Table A1 provides an overview of the two birth cohorts included in our analyses and their exposure to tests. This shows that while the 2008 birth cohort was tested twice (fall 2016, spring 2017), the 2009 cohort was tested three times (fall 2016, spring 2017 and spring 2018).[5] All of the tests were digital, 45 minutes in duration, conducted on a tablet or computer, and developed by educators experienced in developing mathematics tests for students in primary schools. These tests were designed to capture student knowledge of mathematical concepts contained in the

[2] The results from PISA 2000 marks a turning point in the public debate as the results revealed that Norwegian 15-year-olds performed worse than expected – only at about the average for OECD-countries. This spurred several white papers, including the reports form the so-called "Kvalitetsutvalget" that were the first to recommend the introduction of the national tests in grades 5, 8 and 9 in year 2004 (NOU 2002:10; NOU 2003:16).

[3] School leaders in the intervention schools were allocated an additional teacher person-year in the school years 2016/17 to 2019/20, which they were instructed to use for small group tutoring in mathematics in specified grades.

[4] For more information on the design and impact of the intervention see Bonesrønning et al. (2022).

[5] In the initial 1+1 Project two other cohorts were included – children born in 2010 and 2011. However, both of these cohorts were affected by the Covid-19 pandemic prior to completing the national tests, which may impact their outcomes. In addition, the 2010 cohort was only tested once and national tests for the 2011 cohort are not yet available.

curriculum for grades 2-4, with a focus on numerical understanding and elementary arithmetic. As these tests were aimed at distinguishing between different ability levels, they included exercises that all students would be able to solve and exercises that few students would. As can be seen from the distributions of these test scores (Appendix Figure A1) the tests were successful in generating a wide range of test score performance. They are also informative regarding mathematics attainment insofar as these test scores are highly correlated (approximately 0.60) with the later national tests that these children sat 1 to 2 years later. As this was the first-time students in this age group were exposed to such tests, they were allowed to explore some sample exercises to familiarise themselves with this type of test prior to test-taking. For exercises with a lot of text the students could click on a button to hear it read aloud. This was an important feature due to the children's young age as the aim was to capture mathematics, rather than reading, ability.

Schools were instructed that all students should sit the test, with the same exemption rules as used for the National tests in grade 5. That is exemption was only allowed for students with special needs or with Norwegian as a second language. Appendix Table A2 provides an overview of the share of students sitting the different tests by cohort. The rate of students not sitting the tests appears to have been marginally higher than the equivalent rates for the grade 5 national tests. However, the vast majority ended up sitting the test – ranging from 87% to 89% of all students.

Our second focus is on the screening tests introduced in grade 3. Their introduction followed shortly after the initial development of the national testing system (White paper St.meld. nr. 31, 2007-2008). In 2008, the first mandatory screening tests in reading and mathematics were introduced for grade 2. Further compulsory tests in reading for grades 1 and 3 were introduced in 2009 and 2010, respectively, followed by voluntary tests in mathematics for grades 1 and 3 in 2011 and 2010, respectively, and voluntary tests in English and digital skills in 2011 and 2013, respectively.

These tests were designed to screen for students with low educational attainment, where this was based on a student performing below the first quintile. These tests were paper-based, and schools were instructed not to prepare students as doing so would reduce the diagnostic value of the exam. By design they were not difficult for the average student and provided little information beyond highlighting students in need. Yet, the experience of introducing these tests is instructive with respect to the degree of resistance to testing in Norway.

Since the development of the national testing system there has been an ongoing debate about the extent of testing in Norwegian schools. Most recently, following the national elections in 2021, the newly elected government stated in their platform that they would establish a public commission

to review the national quality assessment system with the objective of reducing the number of tests (Hurdalsplattformen). Even though the report from the public commission is not yet available, as of 2022 there are only two mandatory screening tests left – in mathematics and reading in grade 3.[6]

The Norwegian Education Act does not specify any specific rights for students in the 1st quintile nor are they entitled to additional resources. However, the Norwegian Directorate for Education and Training, which is the executive agency for the Ministry of Education and Research, provide general guidelines for head teachers on how to follow-up students that perform in the 1st quintile. These include informing the parents as well as providing them with information on how the school will help the student. Furthermore, the guidelines state that there can be both measures aimed at individualized instruction within the regular classroom and the possibility of providing these students with targeted "courses" with relevant content. It is also stated that all measures should be planned in coordination between the different actors in the school such as a special education teacher, class teacher and mathematics/and or reading instructor. Screening tests may also help identifying special needs students at an earlier age than would otherwise be the case. The rights for students with learning difficulties are specified in the Education Act, §5-5.[7]

3. Data

3.1. School owner and student level register data

The municipalities that were part of the 1+1 project are ten large Norwegian municipalities. This makes it difficult to compare our treatment schools to the entire set of schools outside 1+1, which includes many regional and rural schools that are likely to be different in a number of ways from urban schools in ways potentially consequential for our analyses. We adopt several strategies to select an appropriate control group. These range from comparing the effect of testing in treatment schools against 1) all other schools in Norway to 2) our preferred strategy where we exclude schools from municipalities that are very different from the ten treated municipalities.

To select relevant school owners in the second strategy, we rely upon a system developed by Statistics Norway that groups Norwegian municipalities in terms of population size, revenue, and a calculation of how expensive it is to provide municipal services within that municipality

[6] The previously mandatory grade 2 test in mathematics was moved to grade 3 and all grade 2 tests were removed. The previously mandatory reading test in grade 1 was made voluntary and there is still a voluntary test in mathematics in grade 1. It is left to the municipalities whether they would like to participate or not.

[7] Unfortunately, there is no individual level information on special education so we cannot measure whether the screening tests led to early identification of such students.

(Kringlebotten and Langørgen, 2020). This results in 15 different municipal groups across the 356 municipalities that currently exist in Norway. The municipalities involved in the 1+1 project are taken from two of these groups which contain only relatively large and urban municipalities.[8] We select the other 51 municipalities from these two groups and use schools in these areas as our comparison group. However, as we demonstrate, a critical point is that our results are in practice not sensitive to the choice of comparison group.

Appendix Table A3 provides a comparison between the treated municipalities and these two alternative comparison groups. Compared to the rest of Norway, the treated municipalities have a relatively large population size, share of children and youths and unemployment rate. The elderly population share and the net operating expenditures on schooling are lower. When we shift to the comparison group consisting of 51 municipality, these differences are less stark. However, population size remains larger in the treated municipalities, while at the same time net operating school expenditures are about the same size on average.

Our primary data is drawn from the Norwegian education registers made available by Statistics Norway. These provide information for the whole population of Norwegian primary school students on national tests in grade 5 (our main outcome measure). Background information includes variables such as the school attended, municipality of residence, gender, country of birth, parental country of birth, parental education level and month and year of birth. Each school is identified through a unique school number. For the analysis of the 1+1 project testing, we limit our sample to schools we can follow throughout our analysis period. We merge information about treatment status for each school ID for the 81 schools that received testing but no additional teachers.[9]

We observe test score data for the population of Norwegian 5th graders from 2014 to 2019 and standardise these to mean 0 and a standard deviation of 1. Appendix Figure A2 provides distributions of these test scores for cohorts of approximately 50,000 students each year. This data is merged with individual and family information from Statistics Norway. School-level information comes from the administrative system *Grunnskolens informasjonssystem* (GSI), which is collected annually and contains the status per October 1 and is reported by the principals. Appendix Table A4 provides descriptive statistics for a selection of these characteristics, along with mean

[8] The municipalities are all grouped into groups 13 and 14. See https://www.ssb.no/klass/klassifikasjoner/112/versjon/1450/koder for more details.

[9] For treated cohorts we have information from the 1+1 project on school attended in grade 3 (2008 cohort) and grade 2 (2009 cohort). For prior cohorts we use information on attended school based on school attended when they sat the national test in numeracy in grade 5. For our treatment group less than 10% is registered at a different school in grade 5 than in grade2/3.

differences between tested and non-tested students in our complete and urban sample, respectively.

3.2. Student survey data

To assess if student well-being, motivation and other aspects of the learning environment are affected by testing, we gathered data from a nationwide student survey which has been collected yearly since 2003/04 by the Directorate for Education and Training (https://www.udir.no/in-english/). The survey is only mandatory to answer in the autumn for grades 7, 10 and 11, but all schools are encouraged to include students throughout grades 5-13. The decision on participation is usually taken at the municipal or school level. By inspecting the data, we find that all the schools that implemented the testing regime participated in the survey in relevant years, implying the absence of selection, at least amongst these schools. However, we drop seven schools from the sample due to mergers in grade 5 among schools with and without an additional teacher in grades 2-4.[10]

Although our data consists of all individual responses, the data collection is done anonymously. This means we are unable to merge the student survey data with other individual register data. However, information on year, municipality, school, grade and gender makes it possible to identify treatment and to assess heterogeneity across gender.

The survey data includes a total of 37 items that cover a wide range of subjects such as well-being, motivation, effort, perception of teacher behaviour and guidance, academic challenges, student participation, bullying and school rules.[11] We consider many of these items to have little relevance when assessing the possible effects of testing. Therefore, we focus on 13 questions regarding student well-being, motivation, student and teacher behaviour, and student-teacher interaction. The students answer these questions on a 1-5 Likert scale. Due to GDPR rules, the Directorate does not have permission to share 5th grade data from 2015 and 2016 with researchers, leaving us with three pre-treatment years (2013, 2014 and 2017) and two post-treatment years (2018 and 2019). The control schools consist of schools in urban municipalities, as described in section 3.1. In addition, we have excluded schools with less than ten respondents, leaving us a sample of between 54,628 and 60,192 observations dependent on the response rates to given questions. Appendix Figure A3 displays summaries of responses to these questions and shows that most students agree or highly agree with the different statements they are asked about in the survey. The

[10] These are Løding in Bodø, Lesterud in Bærum, and Byåsen, Flatåsen, Kolstad, Sjetne and Steindal in Trondheim.
[11] The compulsory questions have been included since the last revision of the Norwegian Pupil Survey in 2012 and have been subjected to both exploratory and partly confirmatory factor analysis (CFA) (Federici, Caspersen, & Wendelborg, 2016; Wendelborg, Roe, & Federici, 2014).

last graph in Appendix Figure A3 shows the distribution for the first principal component based on the 13 survey items. This factor component has a higher variation than each survey item, with a standard deviation equal to 2.13.

4. Empirical Methodology

4.1. 1+1 Testing

Our initial approach to examine the effects of the increased test-taking in early grades introduced by the 1+1 project is by applying an event study design. Our main results come from estimating the model in Eq. (1).

$$(1)\ y_{ist} = \alpha + \sum_{t=2015}^{2019} \beta_t \, Test_s \times Year_t + \boldsymbol{X}_{ist}\gamma + \theta_t + \mu_s + e_{ist}$$

where *y* is the dependent variable, most often test scores for fifth-grade student *i* attending school *s* in year *t* (at autumn). The dummy variable *Test* takes the value of 1 for schools that introduced a test in the lower grades but did not receive the 1+1 treatment (additional teacher resources). $Year_t$ are year dummies and is measured as the year students sit for the fifth-grade national tests. We have chosen 2014 to act as the reference year. The β_t vector represents the treatment effect, where 2018 and 2019 constitute the post-testing period. This approach allows us to directly examine the parallel trend assumption and the potential for heterogenous treatment effects across the two treated cohorts. The latter element is interesting since students born in 2009 were introduced to testing earlier than those born in 2008 (age 7 compared to age 8). In addition, the 2009 cohort had to sit one additional test compared to those born in 2008.

The *X* vector includes the control variables gender, immigration status (1st and 2nd generation, respectively), parental education, birth quartile, and the cohort size at student *i*'s school. Parental education is categorised based on three groups. The reference group is students with maximum parental education equal to unknown or primary education only, while the model includes two dummy variables capturing students with parents holding maximum upper secondary and higher education, respectively. μ_s, and θ_t are school and year fixed effects, respectively, α is a constant term and *e* is the error term.

A concern is our choice of control groups due to differences that remain between our treatment schools and the other large municipalities from which our preferred comparison group are drawn. While not inherently a problem for our event study design, an issue remains that there may exist time-varying differences in the evolution of test scores across these schools that differ between treatment and comparison municipalities. A first approach to address this concern is to apply a

semiparametric difference in differences estimator as proposed by Abadie (2005). The estimator represents a generalisation of the conventional difference in differences model in the case when observable characteristics explain differences in the trends of the dependent variable in the treatment and control groups. The estimator adjusts the distribution of the covariates between treated and non-treated units using propensity score matching, see Abadie and Cattaneo (2018). The covariates and the pre-testing outcomes are based on school averages for 2016-2017, whereas the post-testing outcomes are averages for 2018-2019. The analysis unit is, thus, at the school level, where the dependent variable is the difference in outcome between the post- and the pre-testing period.

While we focus initially on the effect of early testing on average educational attainment, there may be effects at other margins. Indeed, much of the argument for early testing emphasises the ability to help students with low educational attainment. We examine this further by using the information on mathematic skill levels of students.[12] These are thresholds set by the Norwegian Directorate of Education and Training to group students into three categories based on test scores on the national tests in fifth grade. Skill level 1 in numeracy suggests that students can only solve simple numeracy tasks in non-complex contexts with a limited amount of strategies. Skill level 2 indicates the potential to solve and reflect upon somewhat more complex problems in non-complex contexts utilising different strategies while skill level 3 suggests that a student can interpret and solve complex numeracy tasks regardless of context. We estimate variants of (1) where instead of continuous test score achievement, our dependent variable is whether the student has reached a given level of skills.

4.2. Screening Tests

The introduction of mathematics screening tests in 2008 simultaneously affected all grade 2 students in Norway (born in 2000). For most of Norway's municipalities, this marks a sharp change in the use of screening tests in mathematics in primary schools.[13] Oslo, Norway's largest municipality covering about 10 percent of the primary school population, is an exception as it developed its own testing regime before the national expansion.[14] At the same time as the screening

[12] See https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/mestringsbeskrivelser-for-nasjonale-prover-i-regning/ for a Norwegian description on the skill levels.

[13] Prior to the introduction in 2008, the test was piloted at 24 schools spread across the country. Ideally, we would like to exclude these schools from our sample, but we do not have identifiable school information in our data. However, given that the pilot only affected 24 out of about 3000 primary schools in Norway, this is unlikely to have any impact on our results. For more information on the initial screening test in mathematics, see Alseth, Throndsen & Turmo (2007).

[14] Among other things, Oslo introduced its own mandatory tests in mathematics for all grade 3 and 5 students in the year 2006 and reading tests for students in grades 2, 5, 7 and 9 in 2007. About 10 percent of primary school students

test in mathematics was introduced, a mandatory screening test in reading was introduced in grade 2. However, this likely crowded out the use of pre-existing standardized voluntary screening tests in reading (see Solheim and Tønnesen, 1998; Waglermo et al., 2018). Hence, the introduction of screening tests in grade 2 primarily provides a change in the use of screening tests in mathematics.

The nature of the introduction creates a sharp discontinuity in treatment by month of birth. Individuals born on or after January 1, 2000, are exposed to the screening test in grade 2. Those born on or before December 31, 1999, were not exposed to screening tests before sitting the national test in grade 5. Our identification strategy exploits this sharp discontinuity in treatment between individuals born a short time apart. The idea is that individuals born just before January 1, 2000, are a good comparison group for those born at or just after January 1, 2000. However, comparing individuals born around the January cut-off may confound the treatment effect with age at school entry/age at test-taking effects (see e.g. Dee and Sievertsen, 2014). January-born children are one year older when they start school and one year older when tested in grade 5 compared to those born in December, as the school cut-off is January 1.

This leads us to use a difference-in-discontinuity design where we use individuals born in the same months but in years where there were no changes in test exposure around the January cut-off. Specifically, we compare results on the national test in 5th grade between children born in January-June 2000 and July-December 1999 (Reform window) to children born in the same two periods in years not affected by changes in test exposure (Control window). For this purpose, we use individuals born in January-June 1999 and July-December 1998.[15] While again we estimate average test score effects of exposure to these tests, skill levels are likely more salient since the screening is only directed at low-performing children.

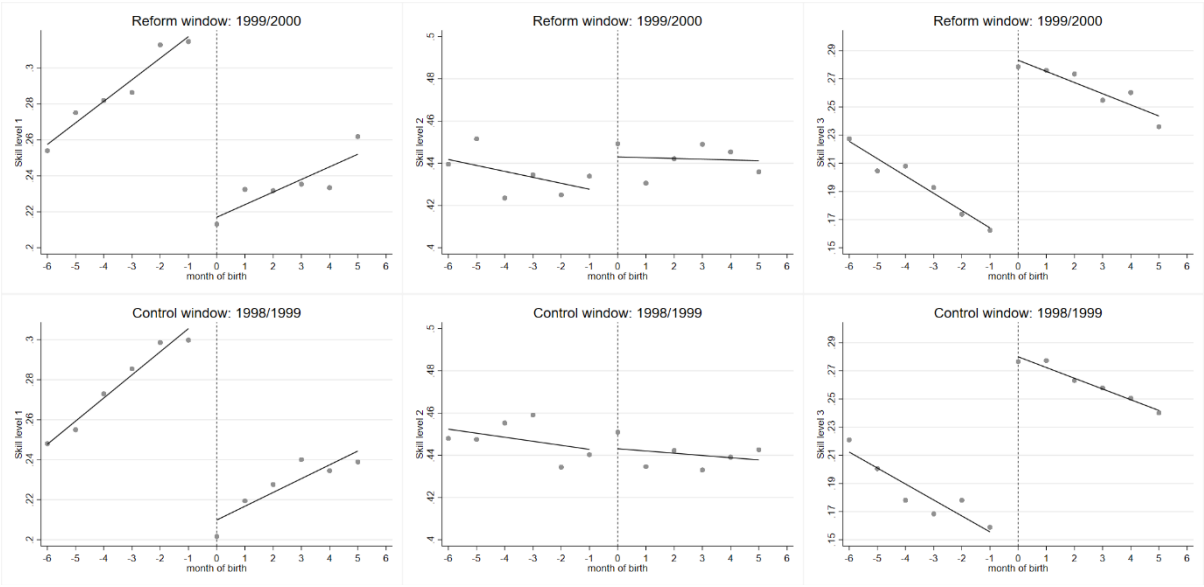
Figure 1 provides a visualization of the school entry/age at test-taking effects by showing the share of students in each skill group of the national test in grade 5 by relative month of birth (relative to January). We show the share of students in each skill group by month of birth (dots) and linear fits on each side of the January cut-off. The regression lines illustrate the jump around the January 1 cut-off. This is shown separately for the Reform window (1999/2000) and Control window (1998/1999). First, consider the Control window. There is a clear relationship between birth month and the share of students in the lowest tertile of the national test (skill level 1), made

reside in Oslo. For further information, see https://www.utdanningsnytt.no/grunnskole/omfattende-testing-i-oslo-skolen/120703.

[15] Similar approaches have previously been used in multiple papers studying the impact of parental leave (Lalive & Zweimuller, 2009; Lalive et al., 2014; Dustmann & Schonberg, 2012; Cools et al., 2015) and the impact of school reforms (Bertrand et al., 2021; Canaan, 2020).

particularly clear by the fall in the share of students in this group around the cut-off. With respect to the second tertile (skill level 2), there is no clear relationship between the month of birth and the share of students in this group. For the third tertile (skill level 3), there is a clear negative relationship between birth month and the share of students in this skill group, made particularly clear by the jump around the cut-off. Moving on to the same figures for the reform window, we see a similar pattern for skill level 1, although the difference between those born in January and December appears somewhat smaller than for the control window. Furthermore, for the second tertile, there appears to be a jump from December to January, a pattern we do not observe in the Control window. Combined, this suggests that the exposure to the grade 2 screening test affected the exposed cohorts' results on the national test in grade 5.

Figure 1. Share of students at each skill level by month of birth



Our empirical model can be written as:

$$(2)\quad Y_{im} = \beta_0 + \beta_1 after_{im} + \beta_2 cohort2000/1999_{im} + \beta_3 after_{im} \times cohort2000/1999_{im} + \gamma_m + \varepsilon_{im}$$

where Y_{im} denotes the outcome of interest for individual i born in month m. $after_{im}$ is a dummy equal to one if the individual was born between January and June, and $cohort2000/1999_{im}$ is a dummy variable equal to one if the individual was born in the period July 1999 to June 2000 and 0 otherwise. γ_m is birth month fixed effects. Under the assumption that age at test-taking and school entry effects are stable across cohorts, β_3 identifies the effect of exposure to the grade 2 screening tests in mathematics.

We estimate this model on a sample of students born in the reform and control windows registered as residing in Norway the year they turn 6 (enter grade 1). Further, we exclude individuals living in

the largest municipality, Oslo, as they were exposed to different mathematics tests earlier. Appendix Table A5 and A6 present descriptive statistics and a balance test, where we estimate the model specified in (2) using different baseline characteristics as the outcome measure.

We also use this strategy as an alternative strategy to measure the impact of the increased test exposure due to the 1+1 project test. We then limit the sample to include individuals in the control group schools of the 1+1 project as the project created a discontinuity in treatment across birth cohorts in these schools. Individuals in affected schools born on or after January 1, 2008, are exposed to more testing, whereas individuals born on or before December 31, 2007, in the same schools are not exposed.[16] Specifically, we compare outcomes on the national test in 5th grade between children born in January-June 2008 and July-December 2007 to children born in the same two periods in years unaffected by test-taking. We use January-June 2007 and July-December 2006 (control group 1), January-June 2006 and July-December 2005 (control group 2) and January-June 2005 and July-December 2004 (control group 3) to control for age at test-taking and school entry effects.

5. Results

5.1. The Effect of Additional Testing on Mathematics Attainment

Figure 2 provides initial treatment estimates of introducing the 1+1 project testing on mathematics performance in grade 5 national tests. The full model is reported in Appendix Table A7. The key take-away from these estimates is that the effect of testing is zero. The point estimates for 2018 and 2019 have small negative signs (not statistically significant) compared to the reference year. The point estimate for 2018-2019 pooled is also negative in a standard difference in differences approach, where the pre-period consists of 2014-2017.[17] The point estimates for 2014 to 2017 are Abetween student characteristics and math performance fit with expectations.[18]

The second cohort received one additional test, and their first test occurred at an earlier age. If testing affects later performance, one might expect these effects to be more pronounced / evident for the earlier and more intensely treated cohort. The estimated effect for the 2009 cohort (national

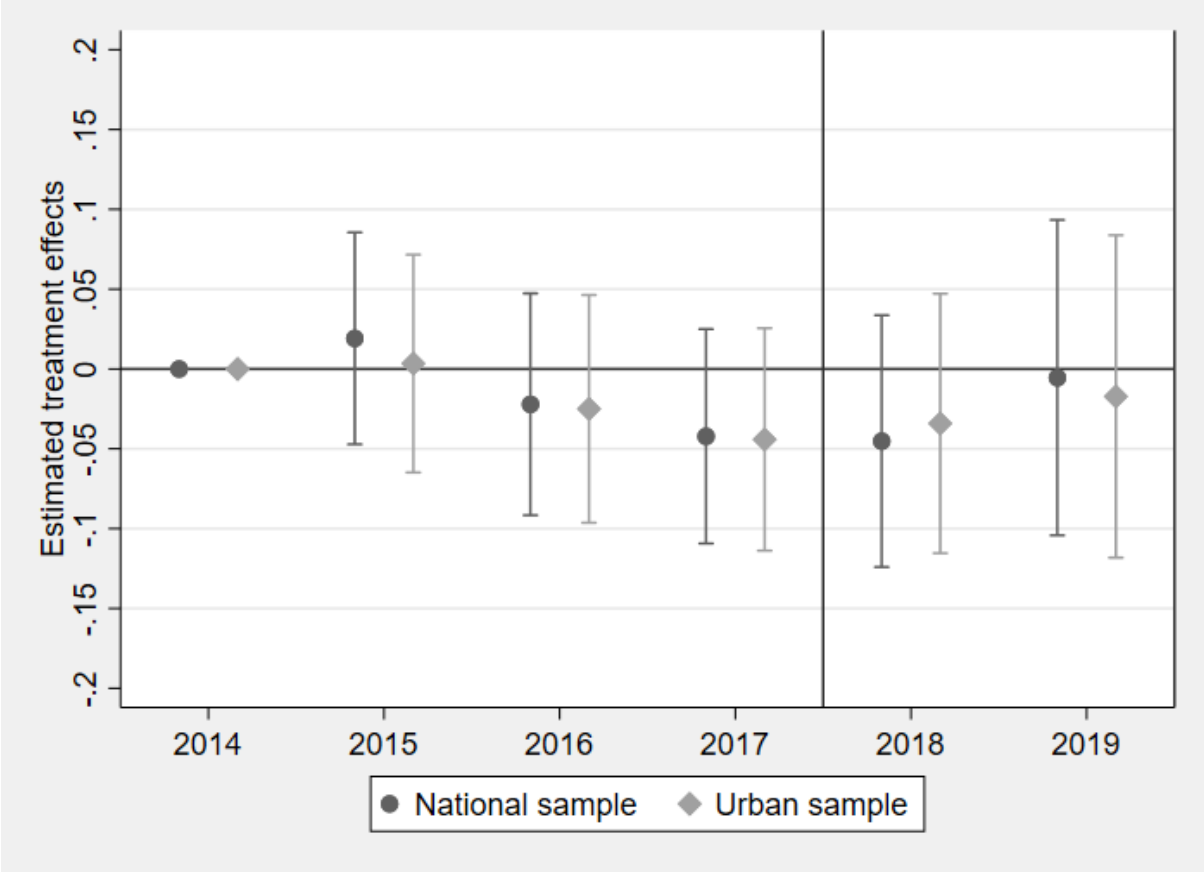
[16] There are 81 treatment schools that we can follow across time.

[17] The model is not reported, but is available on request. The point estimate is -.013 and -.0085 for the national and urban control group sample, respectively. In a simple OLS without fixed effects, the negative point estimates are somewhat higher but far from statistically significant. This analysis also shows that the raw difference between the treated schools and schools in the rest of Norway is .085 of a standard deviation, although much smaller and not significant when including control variables.

[18] Note that there is also no evidence of effects of maths testing on other domains of education performance. Appendix Table A8 reports DiD estimates for both Norwegian and English national tests and demonstrates null results.

test in 2019) is somewhat larger than the estimate for the 2008 cohort (tested in 2018). As shown in Figure 2, in practice, not only do the results follow those presented earlier, the effect of testing is zero for the more intensely treated cohort. This is further suggestive of a true zero effect of early math testing on later math attainment.

Figure 2. Treatment effects on test scores (z-score), national test in numeracy



Note: The full model is reported in Appendix Table A7.

We next adopt two alternative approaches aimed at exploring the robustness of our main results to the choice of comparison group. First, and as reported in Table 1, we adopt a range of semi-parametric matching difference in differences approaches. In the first two columns, we use the full sample of potential comparison schools but first match schools on the basis of pre-treatment period student characteristics at the school level. We then extend this by additionally using average math test scores at a school level as part of the matching approach. The resultant estimates are difference in differences estimates and display clearly zero results of testing. Columns (3) and (4) report similar exercises, but where we first exclude all non-urban schools. Again, this results in an estimate of zero.

Table 1. Testing and numeracy test scores: Semi-parametric diff-in-diff estimates

	(1)	(2)	(3)	(4)
Control group:	National sample		Urban sample	
Match on X-vars	Pre covariates	Pre covariates + test scores	Pre covariates	Pre covariates + test scores
Testing	0.0007	-0.0122	0.0015	-0.0014
	(0.02599)	(0.02403)	(0.02786)	(0.02588)
Observations	1,236	1,226	682	681

Note: Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Next, we adopt the within-school across cohort difference-in-discontinuity approach, as presented in detail in Section 4.2. The results are reported in Table 2 (corresponding balance tests are reported in Appendix Table A9). We report a range of estimates reflecting the different control groups discussed earlier. While these are sometimes less precise, the effects are not statistically different from zero. Together, these approaches further suggest that our main results do not reflect issues in the chosen comparison schools.

Table 2. Testing and numeracy test scores: Difference-in-discontinuity estimates

	Coef	**N**
Panel A: Pooled control group		
*treat*after*	0.003 (0.05)	15,675
Panel B: Control group 1		
*treat*after*	0.007 (0.07)	7,906
Panel C: Control group 2		
*treat*after*	0.035 (0.05)	7,893
Panel D: Control group 3		
*treat*after*	-0.034 (0.06)	7,692

Note: All regressions come from separate difference-in-discontinuity regressions where we include indicator variables for birth month with January/December as the reference category. We include controls for gender, parental education level, immigrant background and cohort size at the school. Standard errors, clustered at the school level, reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

In Table 3, we evaluate the potential effects of early testing along other margins than average educational attainment. More precisely, the dependent variables in these analyses are dummy variables indicating skill level in the national numeracy test. The results do not indicate that the share of students in the three different skill levels has been affected by the introduction of early low-stake testing. These findings are unaffected by utilising a difference-in-discontinuity approach (not reported).

Table 3. DiD effects on mathematics skill levels from the 1+1 project testing

	(1)	(2)	(3)
Skill level	Level 1	Level 2	Level 3
Testing x POST	0.008	-0.008	0.000
	(0.0100)	(0.0074)	(0.0113)
Students in sample	Urban	Urban	Urban
Control variables	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
School FE	Yes	Yes	Yes
Observations	193,188	193,188	193,188

Note: Dependent variables are dummy variables for the three different skill levels, respectively. Skill level 1 indicates low mathematics competence. The models include a constant term (not reported). Control variables included are the same as in Appendix Table A7 (gender, immigration status, parental education, birth quartile and cohort size). Standard errors clustered at the school level are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

A focus of both the previous literature, and the discussion around testing, is that the advantages (and disadvantages) of testing early may be focused within specific groups. In particular, one often emphasised aim of early testing is to identify individuals with additional learning needs so as to target them for intervention and / or additional resources. This is a point we return to later with respect to the analysis of early screening test introduction. Here, we examine this by providing separate treatment effect estimates for a range of groups often thought to be, on average, more educationally disadvantaged.

This leads to estimates in Table 4, where we allow for further interactions between treatment and gender, immigration status, parental educational level, birth quarter and school/cohort size. While these continue to demonstrate zero effects in general, they suggest that the overall zero effects present some heterogeneity across gender and parental background. With respect to the latter, these results are suggestive of a negative effect on children with low educated parents.

Table 4. DiD estimates and heterogeneity, 1+1 testing

	(1)	(2)	(3)	(4)	(5)
Testing x POST	-0.047	-0.009	0.010	0.012	0.020
	(0.0353)	(0.0317)	(0.0314)	(0.0331)	(0.0620)
Testing x POST x Boy	0.074***				
	(0.0271)				
Testing x POST x first gen. immigrant		-0.003			
		(0.0564)			
Testing x POST x Low parental education			-0.090***		
			(0.0327)		
Testing x POST x Born Q3 or Q4				-0.041	
				(0.0282)	
Testing x POST x >P50 cohort size					-0.032
					(0.0705)
Observations	193,188	193,188	193,188	193,188	193,188

Note: Dependent variable is a z-score of test scores in mathematics and the control schools consist of schools in urban municipalities. The models include a constant term, year and school fixed effects (not reported). The control variables included are the same as in Appendix Table A7 (gender, immigration status, parental education, birth quartile and cohort size) and an interaction term between "Testing" and the relevant heterogeneity dimension. Standard errors clustered at the school level are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

5.2. Screening for Children with Educational Difficulties

We next explore the impact of introducing a nationwide mandatory screening test for grade 2 in 2008 using a difference in discontinuity approach as discussed in section 4.2. We focus initially on the impact on whether student results fall into three different skill groups, comparable across cohorts, where 1 is the lowest skill group and 3 is the highest. As discussed earlier, these seem likely salient outcomes given the nature of the screening tests.

Table 5. The effect of screening tests on numeracy skill levels

	Level 1	Level 2	Level 3
Treat*After	-0.004	0.015**	-0.009
	(0.01)	(0.01)	(0.01)
N	103019	103019	103019

Note: All regressions come from separate difference-in-discontinuity regressions where we include indicator variables for month of birth with January/December as the reference category. We control for parental education, gender and immigrant background. Standard errors, clustered at the school level, reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Table 5 presents the estimates. They demonstrate that the introduction of the screening test appears to have increased the share of students in skill group 2 by 1.5 percentage points. We

interpret this as indicating that early screening tests have improved the mathematics performance of some low-performing students.

Table 6. Grade 2 screening test in mathematics, subgroup analysis

	Level 1	Level 2	Level 3
Panel A: Parental education level			
treat*after	-0.003	0.016**	-0.010
	(0.008)	(0.007)	(0.008)
treat*after*LowEdu	-0.003	-0.009	0.003
	(0.019)	(0.019)	(0.012)
treat*after+treat*after*LowEdu	-0.006	0.008	-0.007
	(0.019)	(0.018)	(0.011)
N	103,019	103,019	103,019
Panel B: Minority background			
treat*after	-0.002	0.013*	-0.008
	(0.007)	(0.007)	(0.007)
treat*after*minority	-0.048	0.065*	-0.014
	(0.033)	(0.035)	(0.027)
treat*after+treat*after*minority	-0.050	0.077***	-0.023
	(0.032)	(0.034)	(0.026)
N	103,019	103,019	103,019
Panel C: Gender			
treat*after	-0.017*	0.027***	-0.006
	(0.010)	(0.009)	(0.009)
treat*after*Boy	0.026**	-0.023*	-0.005
	(0.011)	(0.012)	(0.010)
treat*after+treat*after*Boy	0.009	0.004	-0.011
	(0.009)	(0.009)	(0.009)
N	103,019	103,019	103,019

Notes: All regressions come from separate difference-in-discontinuity regressions where we include indicator variables for month of birth with January/December as the reference category. Robust standard errors clustered at school level are reported in parentheses. LowEdu is an indicator equal to 1 if parental education level corresponds to lower secondary school or less. Minority student is an indicator variable equal to 1 if an individual is born abroad. Boy is equal to 1 for boys. treat*after+treat*after*LowEdu, treat*after+treat*after*minority and treat*after+treat*after*Boy is estimated using lincom. *** 1 %, ** 5 %, * 10 %.

Again, this might hide differential effects, especially amongst groups more likely to face additional educational needs. In Table 6, we report the impact for different subgroups – parental education, minority background and gender.[19] It does not appear to be any significant differences by parental education. Panel B investigate whether there are different impact by minority background. The screening test seems to have reduced the share of students performing in the lowest skill group (skill group 1) by 5 percentage points for minority students, whereas this coefficient corresponds

[19] We have checked whether screening test impacted the share of students being exempt from sitting the test, as early identification of struggling students could have increased the share of students with special needs, that are typically exempt from sitting the national test. Our results show no impact on average or by the different subgroups. Results are available upon request.

to 0.2 percentage points for majority students. However, neither the difference nor the coefficient for minority students is statistically significant. Moving on to skill group 2, there appears to be a significantly larger impact on the share of students in this group for minority students. Panel B report differential effects by gender – suggesting that the screening test mainly appears to have moved girls from the lowest skill group to the middle group (skill group 2).

Overall, this provides indicative evidence suggesting that mapping tests mainly move minority students and girls from the lowest skill group to the middle skill group. This could have come about as they were being identified early on to be in the lowest quartile, thereby receiving additional tutoring or more individualized instruction in the classroom.

5.3. Student welfare

A contention is that irrespective of any effects of early testing, testing is harmful to young children and has deleterious effects on teacher and school behaviour. We seek to explore these issues using the combination of student self-reports in the national student learning environment survey (grade 5) and variation in testing through the 1+1 intervention. Hence, we estimate analogous models to (1) where instead our outcomes of interest are student response on questions related to well-being, motivation and interaction with their teachers.

The results are summarised in Figure 3.[20] We report these results for all students and split by gender. A few points are worth making from this. First, increased testing has no effect on whether students report being satisfied at school, are interested in learning at school, or like schoolwork. While there is some suggestion of positive effects on looking forward to school. In summary, there is no evidence of negative effects of testing on student well-being, and if anything, some positive effects.

Looking at interactions with teachers, there is a zero effect on whether (students believe) that teachers care, are respectful, believe in students' learning potential, or provide useful help when needed. However, there is some suggestion that testing improves teachers' ability to 'help students to understand' material. This is also true for the perceived quality of teacher feedback. Again, in summary there is no evidence of a negative effect on (student perceptions) of student-teacher interactions. Where there exist effects, these are positive. There are two questions which cover student ability to undertake work. These are both positive, with the ability to do tasks in class positive. A concern in general is that these results are imprecise and difficult to interpret in their

[20] See Appendix Table A11 for the full regression results for the pooled sample.

totality. Finally, we constructed a principal component of which we report the effects of testing on the 1st component. The estimates from this are positive but not statistically significant.

Figure 3. Effects of testing on student well-being and the learning environment

Note: The figure shows the DiD treatment effects (urban sample) for three models where *1*) all students are pooled, *2*) boys are analysed separately, and *3*) girls are analysed separately. Included control variables are gender, the year-specific number of respondents at the school, year and school fixed effects and a constant term. The standard errors are clustered at the school level. See details in Appendix Table A10 regarding the pooled sample model.

We are quick to admit that these results are far from definitive. What they do provide is a set of evidence that suggests no deleterious effects from testing on student well-being, teacher-student interactions or classroom environment in general.

Conclusion

Whether we should test young children remains a controversial topic. A key issue is if doing so meaningfully improves educational outcomes, and if so, if this is at the cost of child well-being and detrimental effects on school environments. Our paper adds to the small literature in this area by leveraging two sources of variation in math testing among young children. The first is the introduction of a series of high difficulty, but low stakes, mathematics tests to two cohorts aged 7-9 in 81 Norwegian primary schools. The second is the national introduction of easier, but

arguably higher stakes, screening tests in mathematics also in early primary school. This is done in a setting, Norway, which can be viewed as a no to low testing environment at the time of the introduction of these tests.

We demonstrate that the low stakes, high difficulty, tests had a zero effect on later mathematics performance. This is robust to alternative identification strategies and does not appear to vary meaningfully by age of testing or differences in test dosage. In contrast, exposure to screening tests appears to improve mathematics performance of what could be thought of as the targeted groups. There is an increase in the proportion of students performing above the lowest level in mathematics, and these effects appear to be concentrated amongst traditionally disadvantaged groups. In summary, this evidence suggests that the introduction of testing generates zero to small positive effects on later attainment.

A concern often voiced is that the benefits to test score performance, if any, from introducing testing come at the cost of deterioration in student well-being and the school environment. For the introduction of the high difficulty tests, we are able to explore this. Using data from national school environment surveys, we provide a range of evidence on the effects of the introduction of testing from a zero-testing baseline, evidence that we argue is missing from the current literature. We demonstrate no negative effects on student well-being and some evidence that testing improves aspects of teaching practices and students' perceptions of teacher feedback and engagement. Our interpretation is that there is no evidence that testing, at the margin we examine, worsens student well-being or negatively impacts the school environment.

In summary, our results suggest that introducing tests has the potential to have zero to small positive effects on educational attainment, with stronger evidence of positive effects from targeted screening tests. There is no evidence of a well-being-testing trade-off. Together this suggests that there are likely gains to be made from the introduction of testing of young children in low / no test environments, with no indication of trade-offs.

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies, 72*, 1-19. https://doi.org/10.1111/0034-6527.00321

Abadie, A. & Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annual Review of Economics 10*, 465-503. https://doi.org/10.1146/annurev-economics-080217-053402

Alseth, B., Throndsen, I., & Turmo, A. (2007). Rapport fra utvikling og pilotering av "Regneprøven". *Acta Didactica*, *2/2007*.

Anderson, S. C. & Nielsen, H. S. (2020). Learning from Performance Information, *Journal of Public Administration Research and Theory* 30(3), 415–431, https://doi.org/10.1093/jopart/muz036

Bergbauer A. B., Hanushek, E. A. & Woessmann, L. (2021). Testing, *Journal of Human Resources.* https://doi.org/10.3368/jhr.0520-10886R1

Bertrand, M., Mogstad, M., & Mountjoy, J. (2021). Improving educational pathways to social mobility: evidence from Norway's reform 94. *Journal of Labor Economics, 39*(4), 965-1010. https://doi.org/10.1086/713009

Bonesrønning, H., Finseraas, H., Hardoy, I., Vaag Iversen, J. M., Nyhus, O. H., Opheim, V., Salvanes, K. V., Sandsør, A. M. J. & Schone, P. (2022). Small Group Instruction to Improve Student Performance in Mathematics in Early Grades: Results from a Randomized Field Experiment. *Forthcoming, Journal of Public Economics.*

Burgess, S., Wilson, D., & Worth, J. (2013). A natural experiment in school accountability: The impact of school performance information on pupil progress. *Journal of Public Economics*, *106*, 57-67. https://doi.org/10.1016/j.jpubeco.2013.06.005

Canaan, S. (2020). The long-run effects of reducing early school tracking. *Journal of Public Economics, 187*, 104206. https://doi.org/10.1016/j.jpubeco.2020.104206

Cools, S., Fiva, J. H., & Kirkebøen, L. J. (2015). Causal effects of paternity leave on children and parents. *The Scandinavian Journal of Economics, 117*(3), 801-828. https://doi.org/10.1111/sjoe.12113

Dustmann, C., & Schönberg, U. (2012). Expansions in maternity leave coverage and children's long-term outcomes. *American Economic Journal: Applied Economics, 4*(3), 190-224. https://doi.org/10.2139/ssrn.1214910

Dee, T. S., & Jacobs, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(2), 418-446. https://doi.org/10.1002/pam.20586

Dee, T. S., & Sievertsen, H. H. (2014). School Starting Age and Non-Cognitive Skills, *From Birth to Graduation*, 43.

Federici, R. A., Caspersen. J., & Wendelborg, C. (2016). Students' perceptions of teacher support, numeracy, and assessment for learning: Relations with motivational responses and mastery experiences. *International Education Studies, 9*(10), 1-15. https://doi.org/10.5539/ies.v9n10p1

Hurdalsplattformen, 2021-2025: https://www.regjeringen.no/contentassets/cb0adb6c6fee428caa81bd5b339501b0/no/pdfs/hurdalsplattformen.pdf

Högberg, B., & Horn, D. (2022). National High-Stakes Testing, Gender, and School Stress in Europe: A Difference-in-Differences Analysis, *European Sociological Review*. https://doi.org/10.1093/esr/jcac009

Kringlebotten, M. & Langørgen, A. (2020). Gruppering av kommuner etter folkemengde og økonomiske rammebetingelser 2020, Rapporter 2020/48. Statistics Norway.

Lalive, R., Schlosser, A., Steinhauer, A., & Zweimüller, J. (2014). Parental leave and mothers' careers: The relative importance of job protection and cash benefits. *Review of Economic Studies*, 81(1), 219-265. https://doi.org/10.1093/restud/rdt028

Lalive, R., & Zweimüller, J. (2009). How does parental leave affect fertility and return to work? Evidence from two natural experiments. *The Quarterly Journal of Economics*, 124(3), 1363-1402. https://doi.org/10.1162/qjec.2009.124.3.1363

Murphy, R. J.R. Stansfield and G. Wyness (2022) "Teaching to the Test: The Long Run Impacts of Standardised Testing on Student Outcomes" mimeo.

Roediger, H. L. III, Putnam, A. L. & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. Ch. 1. In *Psychology of learning and motivation: Cognition in education*, eds. Jose P. Mestre and Brian H. Ross. Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Solheim & Tønnessen (1998). Kartlegging av leseferdighet og lesevaner på 2. klassetrinn: https://www.regjeringen.no/no/dokumentarkiv/Regjeringen-Bondevik-I/andre-dokumenter/kd/2000/kartlegging-av-leseferdighet-og-lesevane/id105518/

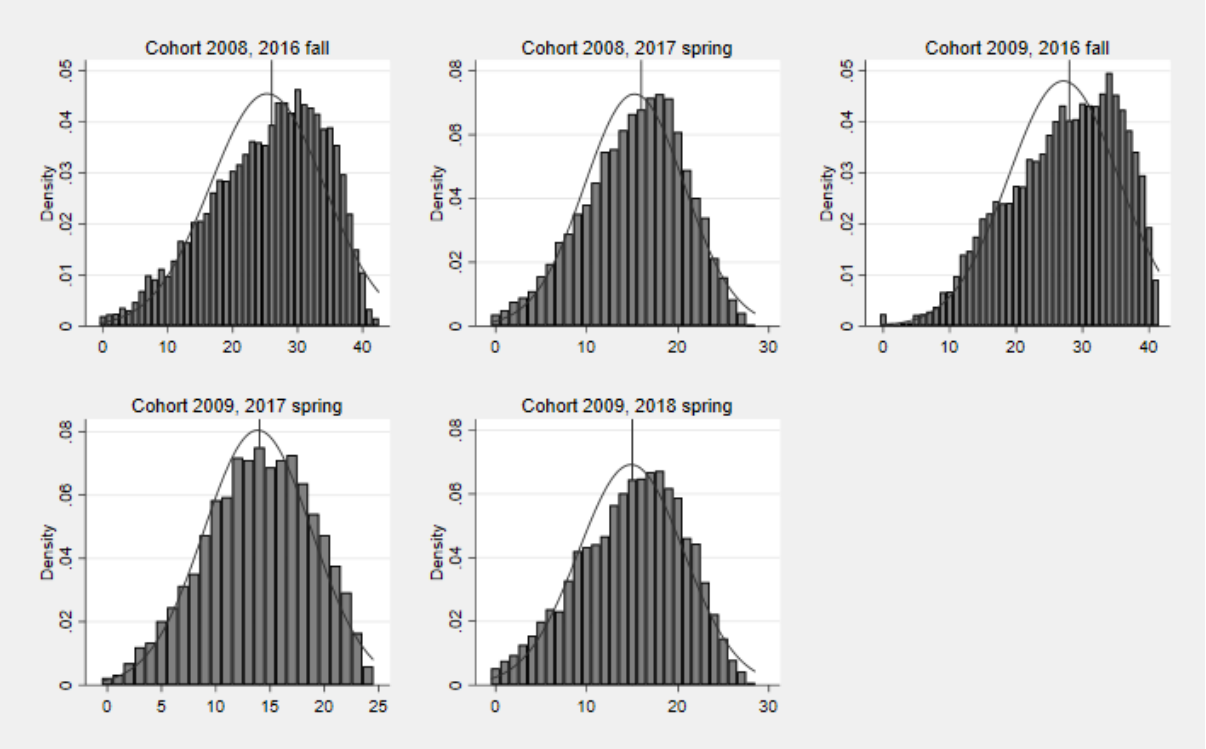
White paper St.meld. nr. 31 (2007-2008): Kvalitet i skolen. https://www.regjeringen.no/contentassets/806ed8f81bef4e03bccd67d16af76979/no/pdfs/stm200720080031000dddpdfs.pdf

Walgermo, B. R., Uppstad, P. H., Lundetræ, K., Tønnessen, F. E., & Solheim, O. J. (2018). Kartleggingsprøver i lesing-tid for nytenking?. *Acta Didactica Norge*, *12*(4), 7-21. https://doi.org/10.5617/adno.6499

Wendelborg, C., Røe, M., & Federici, R. (2014). Elevundersøkelse 2013: Analyse av Elevundersøkelsen 2013. Report. NTNU Social Research.

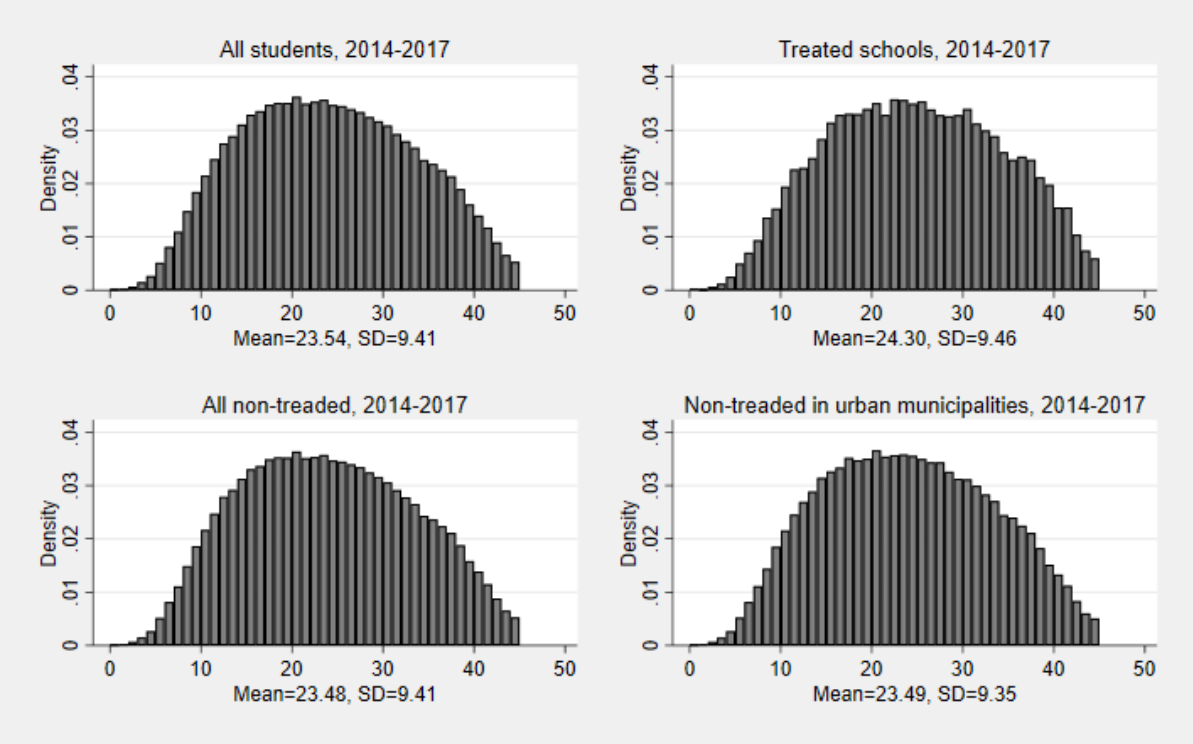
Appendix

Appendix Figure A1. The distribution of test scores in the 159 schools (1+1 Project tests)

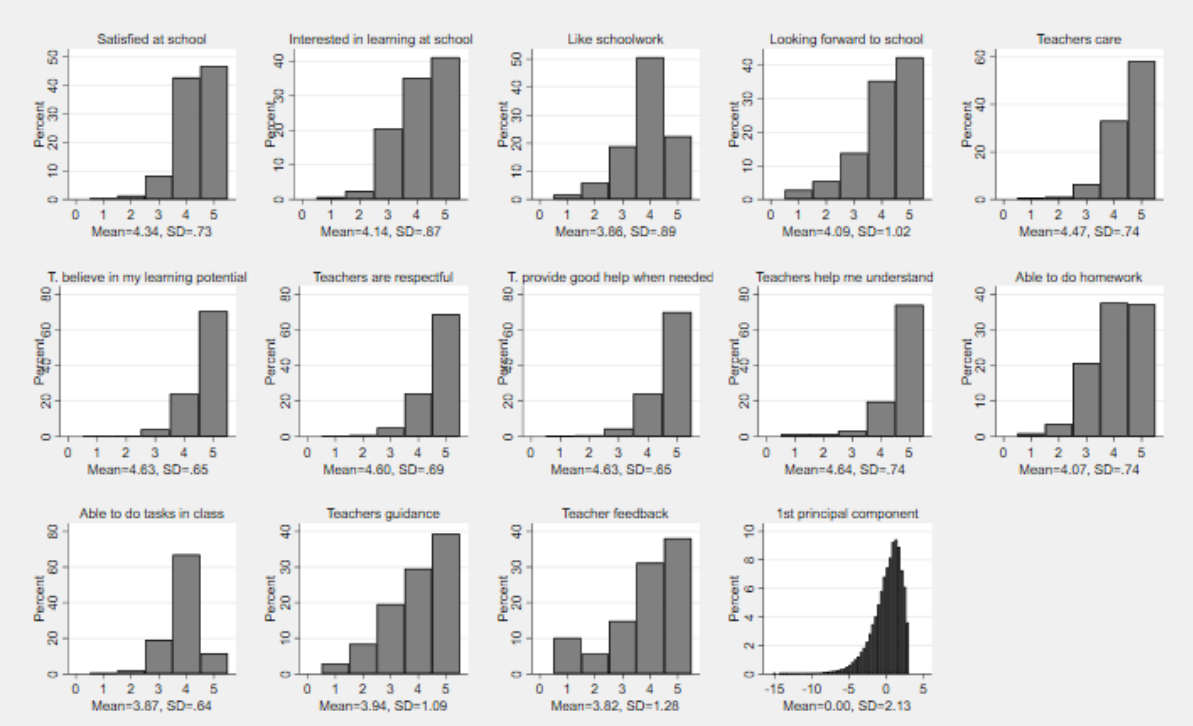


Note: The x-axis measures test score points. The line shows a normal-density plot, while the vertical line indicates the median test score.

Appendix Figure A2. The distribution of test scores for the national test in numeracy, years 2014-2017



Appendix Figure A3. Descriptive statistics on student learning environment survey



Note: The students answer the survey on a 1-5 Likert scale, where 5 indicates high satisfaction/agreement with the statement.

Appendix Table A1: Included cohorts and time of test-taking

	School term		
	Fall 2016	Spring 2017	Spring 2018
Birth year 2008	Grade 3 (age 8)	Grade 3 (age 9)	
Birth year 2009	Grade 2 (age 7)	Grade 2 (age 8)	Grade 3 (age 9)

Note: Age in parenthesis refers to the age of 31.12 in that calendar year.

Appendix Table A2: Percent of students sitting the test

	Fall 2016	Spring 2017	Spring 2018
Birth year 2008	86.8 %	89.4 %	
Birth year 2009	88.3 %	88.8 %	77.7 %

Appendix Table A3. Descriptive statistics on treated and non-treated municipalities

	Treated municipalities	All other municipalities	Urban municipalities
Population	96,072	12,418	39,947
Share children aged 0-5	7.1 %	6.2 %	6.9 %
Share youths aged 6-15	12.3 %	11.9 %	12.5 %
Share elderly aged 67+	14.7 %	19.1 %	16.0 %
Unemployment	2.10 %	1.82 %	2.14 %
Net operating school expenditures per student	96,059	133,974	102,657
Number of municipalities	10	346	51

Note: The descriptive statistics are based on the pre-treatment year 2017.

Appendix Table A4. Descriptive statistics and balance in the pre-treatment period (2014-2017)

	Mean (sd) students in treated schools	Diff. national student sample	Diff. urban student sample[1]
Missing score, numeracy	0.000 (0.000)	-0.001***	-0.000***
Boy	0.512 (.003)	0.001	-0.000
1st generation immigrant	0.089 (.005)	-0.001	0.002
2nd generation immigrant	0.073 (.008)	-0.006	0.001
Low parental edu.	0.079 (.008)	-0.026***	-0.018**
Med. parental edu.	0.217 (.011)	-0.090***	-0.079***
High parental edu.	0.695 (.017)	0.120***	0.098***
Born Q1	0.245 (.003)	-0.002	-0.002
Born Q2	0.262 (.004)	0.002	0.003
Born Q3	0.264 (.004)	0.002	0.001
Born Q4	0.230 (.003)	-0.002	-0.002
Cohort size	57.734 (1.954)	11.780***	10.827***
N treated students		16,105	16,105
N control students		215,918	123,803
Joint F-test		6.540***	5.926***

Note: Standard errors in the joint F-tests are clustered at the school level. The models for joint F-tests also include cohort (year) fixed effects. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

[1]) Municipalities grouped in the KOSTRA groups 13 and 14.

Appendix Table A5. Descriptive statistics, screening test sample

	Mean	SD
Skill group 1	0.255	0.436
Skill group 2	0.441	0.497
Skill group 3	0.228	0.419
Missing test score, numeracy	0.076	0.265
Boy	0.512	0.500
1st gen. immigrant	0.039	0.193
2nd gen. immigrant	0.037	0.190
Parental education = Lower secondary education or less	0.131	0.338
Parental education = Upper secondary education	0.391	0.488
Parental education = Higher education	0.478	0.500

Note: The number of observations is 108,737.

Appendix Table A6. Balance test, screening test

	Coef	SE
Missing test score, numeracy	-0.002	(0.00)
Boy	-0.003	(0.01)
1st gen. immigrant	-0.004**	(0.00)
2nd gen. immigrant	-0.002	(0.00)
Parental education = Lower secondary education or less	0.005	(0.00)
Parental education = Upper secondary education	-0.002	(0.01)
Parental education = Higher education	-0.003	(0.01)
N	103,019	

Note: All regressions come from separate difference-in-discontinuity regressions where we include indicator variables for month of birth with January/December as the reference category. Students in Oslo are excluded from the sample. Standard errors, clustered at the school level, reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Appendix Table A7. Event study estimates

Control group	(1) National sample	(2) Urban sample
Test x Year2015	0.019	0.003
	(0.0338)	(0.0348)
Test x Year2016	-0.022	-0.025
	(0.0354)	(0.0364)
Test x Year2017	-0.042	-0.044
	(0.0342)	(0.0355)
Test x Year2018 (post year)	-0.045	-0.034
	(0.0402)	(0.0414)
Test x Year2019 (post year)	-0.005	-0.017
	(0.0504)	(0.0515)
Boy	0.171***	0.186***
	(0.0037)	(0.0047)
First generation immigrant	-0.223***	-0.218***
	(0.0075)	(0.0097)
Second generation immigrant	-0.217***	-0.216***
	(0.0085)	(0.0101)
Medium parental education	0.232***	0.228***
	(0.0066)	(0.0087)
High parental education	0.654***	0.646***
	(0.0065)	(0.0084)
Born Q2	-0.086***	-0.081***
	(0.0045)	(0.0060)
Born Q3	-0.189***	-0.189***
	(0.0047)	(0.0061)
Born Q4	-0.289***	-0.297***
	(0.0046)	(0.0060)
Students	-0.002***	-0.002***
	(0.0005)	(0.0006)
Constant term	Yes	Yes
Year FE	Yes	Yes
School FE	Yes	Yes
Observations	324,013	193,188

Note: Dependent variable is a z-score of test scores in mathematics. Standard errors clustered at the school level are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Appendix Table A8. DiD treatment effects on Norwegian and English test scores

	(1) Reading	(2) English
Testing x POST	0.0008	-0.0160
	(0.02543)	(0.02848)
P-value parallel trend	0.118	0.023
Students in sample	Urban	Urban
Control variables	Yes	Yes
Year FE	Yes	Yes
School FE	Yes	Yes
Observations	191,230	191,603

Note: Dependent variable is a z-score of test scores in reading and English, respectively. Standard errors clustered at the school level are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Appendix Table A9. Balance test, Difference-in-discontinuity approach

	Pooled control group	**Control group 1**	**Control group 2**	**Control group 3**
Missing score, numeracy	0.004	0.002	0.016*	-0.007
	(0.01)	(0.01)	(0.01)	(0.01)
Boy	-0.006	-0.013	0.001	-0.008
	(0.02)	(0.02)	(0.02)	(0.02)
1st gen. Immigrant	-0.028***	-0.028**	-0.018	-0.038***
	(0.01)	(0.01)	(0.01)	(0.01)
2nd gen. Immigrant	0.002	0.010	-0.001	-0.003
	(0.01)	(0.01)	(0.01)	(0.01)
Cohort size	0.483	0.612	0.844	0.047
	(1.91)	(2.75)	(2.02)	(2.19)
Parental education level :				
Lower secondary	-0.004	0.000	0.000	-0.013
education or less	(0.01)	(0.01)	(0.01)	(0.01)
Upper secondary	0.030*	0.051**	0.018	0.021
Education	(0.02)	(0.02)	(0.02)	(0.02)
Higher education	-0.026	-0.052**	-0.018	-0.008
	(0.02)	(0.02)	(0.02)	(0.02)
N	16,381	8,244	8,264	8,053

Note: All regressions include indicator variables for month of birth with January/December as the reference category. Standard errors, clustered at the school level, are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.

Appendix Table A10. Student welfare, DiD estimates

	(1)	(2)	(3)	(4)	(5)
	Are you satisfied at school?	Are you interested in learning at the school?	How well do you like schoolwork?	Are you looking forward to school?	Teachers care
Testing	0.024	-0.003	-0.012	0.067*	0.019
	(0.0284)	(0.0247)	(0.0289)	(0.0394)	(0.0250)
Observations	60,176	60,192	59,981	59,887	59,874
	(6)	(7)	(8)	(9)	(10)
	Teachers believe in my learning potential	Teachers are respectful	Teachers provide good help when needed	Teachers help to make me understand	I am able to do the homework on my own
Testing	0.009	0.027	0.015	0.028	0.021
	(0.0220)	(0.0213)	(0.0218)	(0.0217)	(0.0274)
Observations	59,670	59,595	59,568	59,570	59,611
	(11)	(12)	(13)	(14)	
	I am able to do my tasks in class	Teachers tell you about your good work?	Teachers tell you what you need to work on?	1st principal component	
Testing	0.030	-0.055	0.083*	0.078	
	(0.0185)	(0.0417)	(0.0475)	(0.0794)	
Observations	57,993	58,834	58,296	53,204	

Note: Dependent variable is a 1-5 Likert variable in columns (1)-(13), whereas it is the continuous first principal component in column (14). See Appendix Figure A3 for descriptive statistics. Standard errors clustered at the school level are reported in parentheses. Coefficients marked ***, **, and * are statistically significant at the 1%, 5% and 10% level, respectively.