The Danish National Tests at a Glance

Louise V. Beuchert [1]
Anne B. Nandrup[2]

This version: September 2017[3]

Abstract: This paper describes the format of the Danish national tests and documents a socioeconomic gradient in student's national test scores. The national test program was implemented in 2010 and covers ten standardized tests in grades 2 through 8. We utilize the complete set of student test scores in 2010-2013 to document a substantial persistence in the children's test scores throughout compulsory school and further show that the predictive validity of the national tests is high in terms of their ninth grade examination results. After establishing a relationship between students' national test results and later education outcomes, the second part of the analysis illustrates evidence on student achievement obtained from the national test data. We document negative and stable test score gaps of considerable magnitude for children with poorer socioeconomic status across all grade levels.

Keywords: Test scores; Student achievement; Test score gap; Predictive validity; Effect size; Education evaluation
JEL codes: I24, I28, J24

1. Corresponding author (lobe@kora.dk); The Danish Centre of Applied Social Science
2. Department of Economics and Business, Aarhus University.
3. The authors appreciate helpful comments from Helena Skyt Nielsen, Simon Calmar Andersen, Maria Knoth Humlum, two anonymous referees, and seminar participants at Aarhus University and TrygFonden's Centre for Child Research. Financial support from TrygFonden's Centre for Child Research is gratefully acknowledged. The authors work with the data from the national tests for research purposes only and are not involved in the development or maintenance of the tests themselves. The usual disclaimers apply.

1. Introduction

The Danish national tests are a comprehensive and mandatory test program in the public primary and lower secondary school. The tests were introduced in 2010 as an integral part of a larger schooling reform responding to the recommendations made by OECD as well as continued disappointing PISA results. A main recommendation in OECD (2004) was to strengthen evaluation in public schools by improving the instruments of teachers for assessment and feedback. The tests are self-scoring in an online, adaptive program. As teachers do not assess the tests, the program ensures that all students are evaluated by the same standards and therefore the results are comparable among all students.

The national tests as an evaluation instrument are twofold. First, at the individual level, the teacher provides individual student feedback based on the test score and integrates the results into individual teaching plans. Teacher feedback is considered one of the main channels to increase individual achievement, see e.g. review by Hattie (2009). Second, in accordance with previous research documenting positive effects of nationwide test programs (see, for example, review by Figlio and Loeb 2011), the average national test results are monitored at the school and national level. A leading concern about test programs is how they may change teacher and student incentives, e.g. by promoting teaching-to-the-test or by demotivating students. Rambøll (2013) evaluated the implementation of the Danish national tests, and found positive impacts of national testing on students' reading and math achievement (although insignificant in math). Andersen and Nielsen (2016) show that the positive impacts are not just attributable to teaching-to-the-test.

This paper provides novel, descriptive evidence on the achievement levels and gaps among Danish students.[4] In particular, the panel structure provided by the national, mandatory tests in reading and math allows us to follow and compare students' achievements across time. Previously, exit exams in ninth grade and international student assessments, such as PISA and TIMMS, provided the only systematic testing of students, but these are cross-sectional only.

We utilize the complete set of test results from 2010-2013 for public school students combined with family and background characteristics, and document considerable discrepancies in student achievement in Denmark: significant gaps in test scores exist between students with a higher and lower socioeconomic status. These gaps are present at all available grade levels (grades 2 to 8). A social gradient in student achievement is not a novel phenomenon, but this paper is the first

4. An earlier version of this paper was circulated as Beuchert and Nandrup (2014): The Danish National Tests: A Practical Guide. *Economics Working Papers* No. 2014-25. Department of Economics, Aarhus University.

to document these on a national scale in Scandinavian compulsory school using standardized, and non-teacher assessed, tests. Sweden, Norway, and Finland have also implemented national test programs, however, their tests are teacher assessed and do not provide reliable, comparable results across students to the same extent as a computer-based and adaptive system with no teacher assessment (OECD 2013). For an example, see Hinnerich and Vlachos (2013; 2017) on Swedish tests.

The empirical analysis is focused on two main questions. First, we consider the persistence and predictive validity of the national test results in terms of test scores and attainment at later stages of education. We show that students' national test scores in early grades are strong predictors for their national test scores in later grades as well as ninth grade exam results and progression to upper secondary education (general and vocational), suggesting that the national tests measure skills that are highly correlated with the skills important for obtaining further education.

Second, the paper provides evidence on student achievement obtained from the national test data. We illustrate how yearly and mandatory testing of students contributes with evidence on socioeconomic disparities among Danish children. We show how student achievement as measured by the national tests is associated with multiple background and parental characteristics such as attained education, income, and immigrant status. Similar OECD (2015), girls' reading skills are on average significantly better compared to boys', however, this pattern is reversed for math based subjects (and English). We then graphically illustrate that gaps in test scores exist and how these gaps develop across age comparing different groups of students. We demonstrate that significant socioeconomic gaps in test scores are present very early in primary school. Furthermore, the gaps seem to persist throughout compulsory school. These insights are important for school policies addressing social mobility at different stages of education.

The remainder of this paper unfolds as follows. Section 2 introduces the national tests and adaptive testing. Section 3 describes the available data and the sample selection followed by the empirical analyses in Section 4. Finally, Section 5 concludes.

2. The National Tests

Following the 2006 schooling reform, all children enrolled in Danish public schools are required to take ten national tests during compulsory schooling; a reading test every second year from the second grade, a math test in grades 3 and

6, and other subject-specific tests in grades 7 and 8.[5] Furthermore, teachers may opt to test students twice in the grade level prior to/after the intended level on a voluntary basis; see Table 1.

Table 1. Grades and subjects tested in the national test program

Subject of the test	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9
Danish, reading		X		X		X		X	
Mathematics			X			X			
English							X		
Geography								X	
Physics/ Chemistry								X	
Biology								X	
Danish as second language[a)]					X		X		

Notes. X indicates the grade levels subject to the national tests with the option of testing students on the grade level above or below (shaded). [a)] The tests for Danish as second language is voluntary.

Each test simultaneously tests three cognitive domains, called profile areas. For example, the reading tests assess language comprehension, decoding, and reading comprehension, while the math tests assess numbers and algebra, geometry, and applied mathematics. For a complete list, see Table A1 in the Appendix. The mandatory tests are completed annually in pre-defined periods, usually in mid-January through April, with a retesting period for absentees in June.[6]

The national tests are IT-based, adaptive, and objective, thus, the students are tested online. Students log on the test website with their unique login. Answers and test results are saved in a personal electronic profile. The subject-specific teacher has access to the electronic profiles for feedback purposes; otherwise, the individual test results are strictly confidential. The test questions are typically multiple choice. Reading test questions involve, for example, word-to-picture matching, word splitting, or reading a text and answering content-related questions. Teachers may aid according to the everyday needs of the student or individually discontinue the test for a student losing focus. The test conditions are de-

5. Besides nationwide testing, the reform includes compulsory exams in ninth grade, individual student teaching plans, and quality assessment reports at the municipal level (Public School Act 2006).
6. Voluntary tests are conducted in October-December. We will not discuss these further.

scribed in the student plan; however, currently we do not have access to this information.

Objectivity arises as teachers are not involved in asking the questions or evaluating the answers. The online test system draws the questions (called items) from a large national item bank and calculates the test results. An advantage of objective and self-scoring tests is that potential teacher bias is reduced. Previous studies have documented teacher bias in the evaluation of student skills, i.e. implying that characteristics, such as gender and race, significantly affect teacher perceptions of student performance (e.g. Dee 2007, Downey and Pribesh 2004, and Rangvid 2015 for Denmark).

To capture the wide range of student proficiency levels in a typical classroom, the national tests are designed as adaptive tests. Simplified, adaptive testing means that the student is presented with items (questions) of varying difficulty based on a continuous assessment of the student's proficiency level. Then, as opposed to regular linear tests, it does not matter how many questions the student are able to answer correctly, instead the difficulty levels of the correctly answered questions are of importance.[7] In short, the adaptive test program draws questions with difficulty levels approximately equal to the proficiency of the student based on his or her history of answers. The psychometric model underlying the adaptive testing algorithm is a Rasch model (Rasch 1960). The Rasch model incorporates a method for ordering individuals according to their skill level, and ordering items according to their difficulty. These are ranked on a continuous and Rasch calibrated logit scale ranging from -7 to 7 (one scale is calibrated for each subject and profile area). The item difficulty level is assessed by a test pilot of approximately 700 students (Rambøll 2013). The final test score, termed *estimated student skill level*, is given on a comparable logit scale on the [-7; 7] interval. For an introduction to the Rasch model, see e.g. Bond and Fox (2007).[8]

Figure 1 illustrates an example of an adaptive test process. Each »×« marks the item given within a single profile area. The light-blue line illustrates the estimated student skill level and the dark-blue line illustrates the difficulty level of the items. The adaptive test process and a large item bank ensures that these closely mirror each other. The first item presented to a student within each cognitive domain (profile area) is designed to have difficulty level 0 (corresponding to around average). As illustrated in Figure 1, the next four items within the same

7. Adaptive test systems are known from international student assessment systems such as the Early Childhood Longitudinal Study (ECLS). For a discussion of student assessment systems, see Jacob and Rothstein (2016).
8. The Rasch model may be viewed a one-parameter version of Item Response Theory (IRT) models. IRT models consider the probability that a student answers each test question (item) correctly as a function of the student's latent ability and characteristics (parameters) of the item. The Rasch model incorporates only one item parameter: the item's difficulty (Jacob and Rothstein 2016). Other IRT models add additional item characteristics

profile area are chosen based on the student's answer to the previous item: A correct (wrong) answer triggers the next item of that profile area to be of a difficulty level of approximately 1 logit above (below) the previous level. Hence, the difficulty level of the test items is very volatile in the beginning of a test period—the student skill level trivially mirrors the item difficulty level in this run-in period. Critics have voiced their concerns that this causes some particularly skittish students to be 'trapped' at too low initial difficulty levels, because a wrong answer is punished relatively harder in the beginning.[9] From the sixth item and onwards the student skill level is iteratively estimated using the Newton-Raphson method and system draws items of a comparable difficulty level; see also Beuchert and Nandrup (2014). Thus, the student is given an item of approximately the same difficulty level as the estimated student skill level based on the sequence of items already answered. This Rasch algorithm implies that a student should be given questions with equal probability of a correct/false answer.

The process continues at least until the skill level estimate satisfies a standard error of measurement (SEM) below 0.55.[10] In short, the SEM denotes the variation in the student's ability to correctly answer the items or the statistical uncertainty of the estimated skills. The SEM is illustrated with the green lines in Figure 1. Like the skill level estimate, the SEM is also (re)estimated after each item answered following the fifth item; see the dark-green line. In the example, the student starts out with a SEM of 1 and reaches an estimated SEM of 0.55 around the 11th test item. By answering additional items the SEM is further reduced and the estimated skill level (light blue) is converging to 1.38. The teacher has no influence on which items the test system draws but can monitor students' test sessions, including the SEM, online and may terminate or prolong the test session. The Ministry of Education recommends continuing testing throughout the booked time slot to ensure the lowest possible uncertainty of the results.

9. From 2015, there are some changes to the test program, including modifications of the run-in period. The run-in period is reduced to three items and the difficulty level of item two and three is adjusted by +0.5 (–0.5) logits following a correct (wrong) answer. Further, the difficulty level of the first item now accommodates the mean difficulty level within specific cognitive profile areas (Undervisningsministeriet 2015). Please consult the Ministry of Education for the latest description of the adaptive algorithm. Beuchert and Nandrup (2014) provides a detailed overview of raised concerns and criticisms regarding the tests.
10. SEM = $1/\sqrt{s^2}$, where s^2 is the sum of the variances of the items that the student has attempted to answer. Originally, a SEM below 0.3 was the limit of sufficiently precisely estimated parameters. In practice, the statistical uncertainty of the skill level estimates is substantially larger (0.55). The estimated skill level has a 5% confidence interval of $\pm 2 \cdot SEM$ (Undervisningsministeriet 2012; 2014).

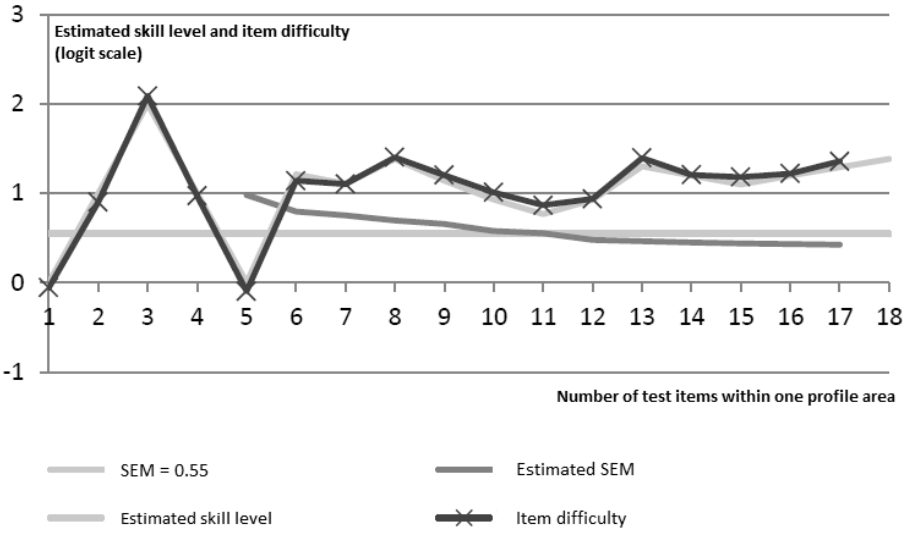

Figure 1. The test process for a single profile area exemplified
The figure illustrates how the estimated student skill level, item difficulty, and estimated standard error of measurement (SEM) may progress during the test of a single profile area (adapted from Undervisningsministeriet 2012).

The resulting test scores are measured continuously and are argued to give a more precise and detailed estimate of a student's skill level compared to what can be revealed by regular linear tests (Review 2007, OECD 2013). To ease interpretation of the test results for teachers, parents, and other stakeholders, they are transformed by a sigmoid (S-shaped) function and reported on a scale from 1-100 points and in one of five norm-referenced groups: considerably below average (1–10 points), below average (11–35 points), average (36–65 points), above average (66–90 points), and considerably above average (91–100 points). The norm is from the 2010 pilot.[11] The norm-based grading reference is a political decision to monitor the achievement level of cohorts across time, for example, to compare the performance of second graders in 2012 to those in 2010. The Public School Act states a set of national goals, one of which is to raise the proportion of students 'considerable above average'.

The teacher has access to test scores reported in 1–100 points and the corresponding norm-referenced group for each student and cognitive domain as well

11. In 2010, 15,000-22,000 students participated in pilot testing of the national tests to evaluate the test properties and calculate a national norm for future reference. This implies that the 1–100 points does not correspond to actual percentiles in a given test year.

as student and class averages. Parents to tested students receive a short letter explaining the results in terms of the norm-based five-group scale only (for an example, see Wandall 2011).

3. Data and sample selection

We sample all students enrolled in public mainstream classrooms in the school years 2009/2010–2012/2013. Using unique, administrative identifiers, we then match students to their parents and other registers on student and school characteristics. We describe these in detail below.

3.1. Background characteristics

We construct a large and detailed panel with yearly information on each student including school and class affiliation and multiple characteristics of the school, the student, and his or her parents. The characteristics of the schools include school size, class size, and indicators for location in capital or larger city area. Parental characteristics include age, marital status, and socioeconomic characteristics such as attained education level and income. Information on family structure and parental characteristics are measured in the year the child turned five, i.e. before school entry, to avoid confounding factors. Student characteristics include gender, birth information (year and quarter of birth, birth weight, and gestational age) and immigration status including origin of birth. We further add information on psychiatric diagnoses (at age 8) and referrals to special needs education (in the previous school year).[12] See a complete list of covariates including sample means in Appendix Table A2.

3.2. Main variables

The main variable of interest is the national test result obtained in grades 2 to 8. To achieve greater precision and avoid interpretational difficulties, we base our analyses on the estimated student skill level on the continuous, Rasch-calibrated logit scale for each cognitive domain. To measure students' overall skills in e.g. reading, we construct a standardized test score measure combining student skill levels in all three cognitive domains. First, we standardize the estimated student skill level to mean zero and unit standard deviation within each year, subject, and cognitive domain to ensure comparability across domains. Then, for each student

12. Psychiatric diagnoses are obtained from hospital registers and include all group-F classifications in WHO's International Classification of Diagnosis (ICD-10 classification in parentheses). We obtain primary cause of special education needs from school registers and define five categories: Physical disabilities (H15-H18), mental disabilities (H11), social disabilities (H12), learning disabilities (H10, H13-H14), and other causes (H20, H99).

and test, we calculate the average across the three cognitive domains and this mean is once again standardized to mean zero and standard deviation one within each year and test. We use this standardized test score as the measure of student achievement within a given test throughout the rest of the paper. The latter standardization allows us to readily interpret regression coefficients as standard deviations.[13]

The national tests are mandatory and the vast majority of the students comply: The average response rate is 85% in 2010, 95% in 2011, 96% in 2012, and 95% in 2013 (see Appendix Table A5 for a detailed overview of the response rates).[14] There are, however, no formal sanctions imposed on schools or truant students. Individual exemptions from test taking are granted only if school representatives, in agreement with the parents, believe that the student is unable to obtain a result that is useful in the evaluation of the child's teaching plan. The fraction of missing test results explained by exemptions is less than 12%. Not surprisingly, students with special education needs have a significantly higher probability of exemption. However, they also have a higher probability of truancy. Being of non-Western background does not increase the likelihood of exemption or missing the national tests once other family and child characteristics are controlled for (all results are available on request).

Additional outcomes of interest include the later educational attainment of students: ninth grade exam results, and enrollment and completion in upper secondary education (general or vocational). The ninth grade exit exams consist of mandatory exams in the subjects Danish (reading, writing, spelling and oral performance), math (calculus and problem solving), English (oral), and physics/chemistry (oral).[15] The grading scale is ordinal with marks -3, 00, 02, 4, 7, 10, and 12, where marks 02 or above pass. Information on exam results are available until 2013, thus, we are able to link student test scores from grades 6 to 8 to their ninth grade exam results for up to three cohorts of students. This linkage is possible for more than 90 percent of all students in each cohort.

13. Table A3 in the Appendix shows that the raw test scores are moderately correlated across cognitive domains within test subjects (between 0.55 and 0.81). These correlations could be caused by other underlying attributes, or simply that the domains themselves overlap. It is beyond the scope of this paper to validate the Rasch properties and reliability of the test items (scoring on individual items are unavailable to researchers). The correlations between the average standardized test score, we construct within each test, and the separate cognitive domains are all above 0.82, see Appendix Table A4.
14. The low response rates in 2010 are caused by a nationwide, technical breakdown that unexpectedly cancelled two full weeks of tests. From 2012 to 2013, the response rates decrease slightly due to a five-week lockout of roughly 80% of the teaching staff in public schools.
15. For Danish and math, we calculate average exam results (GPA). GPAs are adjusted if one or more exam results are missing.

We follow students' enrollment in upper secondary education (general or vocational) after completion of compulsory school (ninth grade), for the three oldest cohorts of our sample. There are four academic-oriented upper secondary education programs: general upper secondary program (STX), higher preparatory program (HF), higher commercial program (HHX), and higher technical program (HTX); all 2- or 3-year programs preparing for higher education. If students wish to pursue a vocational education, more than 100 main vocational education and training programs are available and the duration typically varies from 2 to 4 years.[16] We include all of the above in our measure of enrollment after compulsory school.

4. Empirical Analysis

To date, Danish researchers and policymakers have only had access to predetermined socioeconomic characteristics when predicting achievement of students. It makes an interesting case, if student performance on the national tests can facilitate improved screening for targeted educational policy compared to (or combined with) other observable characteristics such as immigration background or socioeconomic status.

In the first part of the empirical analysis, we consider the predictive validity of the national tests in terms of students' exam results in the ninth grade as well as test results in later grades. Establishing predictive validity is important when discussing national test results as assessment tool for policy. In the second part of the analysis, we illustrate how the national test results can provide new insights about student achievement in schools.

4.1 Predictive validity of the national test results

The national tests are designed to measure student competences in specific cognitive domains through primary and lower secondary school. However, given their early age, little evidence exists on the relation between test results and other measures of later success. Here, we present evidence of the associations between student achievement as measured by the national tests and ninth grade exit exam results. Similar to the national tests, written exit exams are nationwide and standardized. They are graded by teachers (one internal and one externally appointed) and the formal assessment guidelines specifies, in addition to specific course curriculum, broader learning objectives such as the students' understanding and reflection of own learning. As such, exam assessments should reflect the course objectives and general skills valued by teachers. Thus, a high correlation between

16. We condition on enrollment in a vocational main program, thus, we disregard youths enrolling in introductory programs only (6 months).

national test scores and exam results strongly suggest validity of the national tests in terms of measuring a set of skills comparable by those evaluated by the exit exams. Moreover, exam results are generally associated with later success measures, e.g. successfully completing high school or vocational college (see e.g. Hvidtfeldt and Tranæs 2013; Humlum and Jensen 2010).

Examination grades and national test scores may differ for reasons other than differences in measurement scales. First, national test results are self-scoring while the teacher and an external censor grade exams. Second, in case of the oral exams students typically draw only one or two topics from the curriculum to present. Compared to this, the national tests contain items that are relevant for the specific cognitive domain on a more general scale (the questions compose a series of random draws of single items within the respective cognitive domains). Third, as the purpose of the exit exams and the national tests differ, they are likely to measure somewhat different sets of skills. Finally, results from the national tests are low stakes compared to the results from the exit exams, which are qualifying to further education. The test environment of the national tests is considerably more informal compared to that of the exit exams: Students in need are allowed to take breaks and interact (to some degree) with their teacher during the test session. This may cause some students to perform better as exam jitters are less pronounced, while others may perform poorer because stakes are low. Also for schools and teachers, the national tests are low stakes and are not used for sanctioning. The school average adjusted for socioeconomic composition of students are available to the individual school management as well as the municipality in which it resides for comparison with a national adjusted average, which is publicly available.

Table 2 presents the results of regressing students' exam marks on same- and cross-subject national test result obtained in earlier grades. On the individual level, test scores from the national tests alone explain 48-51% of the variation in average Danish and math exam marks. For the oral English and physics exams, the corresponding numbers are 42% and 23%, respectively. In all subjects, the raw estimation results suggest that increasing the test score by 1 SD is associated with approximately 2 grade points' increase in the GPAs.[17] Interestingly, the point estimates on both the sixth and eighth grade reading scores are practically identical. Thus, reading skills in grade 6 seems to be just as strong a predictor of exam results in Danish as reading skills in grade 8, even though two more years of learning have taken place in between. In general, the same-subject correlations are higher than the cross-subject correlations. For example, a higher test result in math is associated with markedly higher math exam result than in the linguistic

17. Controlling for baseline covariates reduces the point estimates; see Appendix Table A6.

subjects Danish and English. Note that better test scores in reading is associated with a considerable improvement of exam results in all four mandatory subjects. Table 3 shows that the correlations are stronger for the low teacher-discretion exams in reading and spelling (approx. 2.1 grade points) and smaller for the high-discretion oral and written essay exams (approx. 1.8 grade points). This is consistent with the three separate cognitive domains of the national tests, which are more closely related to reading and spelling (see Appendix Table A1). It may also reflect similarities of the examination types and the national test process and items.

Table 2. OLS estimates: ninth grade exam marks on national test results

	(1) Danish, GPA		(2) Math, GPA		(3) English, oral		(4) Science, oral	
National test results:								
Reading, grade 6	2.010	***	1.883	***	2.104	***	1.730	***
	(0.013)		(0.015)		(0.019)		(0.019)	
Observations	46,728		46,298		45,435		45,280	
Reading, grade 8	1.997	***	1.988	***	2.213	***	1.825	***
	(0.009)		(0.012)		(0.013)		(0.021)	
Observations	138,970		138,044		136,265		45,476	
Math, grade 6	1.495	***	2.229	***	1.400	***	1.825	***
	(0.016)		(0.018)		(0.021)		(0.021)	
Observations	46,909		46,484		45,616		135,486	
English, grade 7	1.715	***	1.636	***	2.420	***	1.456	***
	(0.010)		(0.012)		(0.015)		(0.014)	
Observations	93,602		92,884		91,331		90,670	
Science, grade 8	1.194	***	1.796	***	1.393	***	1.830	***
	(0.016)		(0.022)		(0.019)		(0.021)	
Observations	138,164		137,308		135,548		134,929	
Mean outcome	6.718		6.659		7.388		6.236	
Covariates	No		No		No		No	

Notes. Each cell reports the estimate from separate regressions of students' exam results (columns) on national test results (rows). All specifications include a constant and year fixed effect. Columns (1)–(2) is based on the average exam marks in Danish (oral, essay, spelling, and reading) and math (problem solving and arithmetics), respectively. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table 3. OLS estimates: ninth grade Danish exam marks on national test results by examination type

	(1)	(2)	(3)	(4)
	Oral exam	Written exams		
	Oral	Essay	Spelling	Reading
National test results:				
Reading, grade 8	1.785 ***	1.869 ***	2.240 ***	2.110 ***
	(0.012)	(0.010)	(0.012)	(0.012)
Observations	137,697	138,150	137,986	138,200
Cohorts	3	3	3	3
Mean outcome	7.456	6.383	6.454	6.416
R-squared	0.212	0.319	0.459	0.416
Covariates	No	No	No	No

Notes. All specifications include a constant and year fixed effect. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table 4 supports the association between students' national test scores and later education attainment: Increasing reading scores by 1 SD is associated with a 18.8 percentage points higher probability of enrolling in upper secondary education (general or vocational) within two years after compulsory school (column (1)). For the oldest cohort of our sample, we can track 4 years after compulsory school and find similar strong correlations on the probability of still being enrolled or having completed (column (2)) and having completed general upper secondary school (column (3)).[18]

[18] Controlling for baseline characteristics, the coefficients decrease by less than two percentage points; see Table A7 in the Appendix.

Table 4. OLS estimates: enrollment and completion of upper secondary education (general or vocational) on national test result

	(1)		(2)		(3)	
	Enrolled 2 years after compulsory school		Completed or enrolled 4 years after compulsory school		Completed general upper secondary 4 years after compulsory school	
National test results:						
Reading, grade 8	0.188	***	0.108	***	0.163	***
	(0.001)		(0.002)		(0.003)	
Observations	141,558		44,073		44,073	
Cohorts	3		1		1	
Mean outcome	0.776		0.830		0.361	
R-squared	0.181		0.076		0.105	
Covariates	No		No		No	

Notes. All specifications include a constant and year fixed effect. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Next, we focus on the persistency and predictive ability of the national test results obtained at earlier grade levels. In particular, we examine how national test scores from previous grades predict later national test scores. We find a remarkable persistency in the test scores across time suggesting that a student's test scores already in early grades provide indicative information on how the student will fare in later grades. Generally, increasing previous achievement by 1 SD improves student achievement by 0.6–0.7 SD, see Table 5.[19] In large samples, it is very unlikely to find these stable correlations if the relationship between same-subject test scores across grades were caused only by noise or chance. From this, we infer that the national tests validly measure the same set of skills across grades. The R-squared values are relatively high; previous reading scores explain 50% or more of the variation in the current test scores, previous math scores slightly less.[20]

In line with the reasoning that reading is an important prerequisite to learn in other subjects students' reading scores predict math scores in later grades to a larger extent than math scores predict later reading scores (results are available on request). However, a considerable part of the correlation is likely to reflect a general interest in learning.

19. 1 SD compares to moving the median-student from 50 points to 84 points, or a low-performing student from 10 points to 39 points.
20. The relationship is as good as unchanged when student and family characteristics are included and the R-squares increase only slightly, see Table A8 in the Appendix. Thus, student baseline characteristics add very little explanatory power to our model—the previous test score, obtained at an earlier grade level, already catch most.

Table 5. OLS estimates: National test results explained by previous test result in the same subject

	(1) Reading, grade 4	(2) Reading, grade 6	(3) Reading, grade 8	(4) Math, grade 6
Previous national test results:				
Reading, (grade -2)	0.686 ***	0.760 ***	0.744 ***	
	(0.004)	(0.004)	(0.004)	
Math, grade 3				0.592 ***
				(0.007)
Observations	90.194	92.922	87.110	43.827
Cohorts	2	2	2	1
Mean outcome	0.058	0.067	0.089	0.067
R-squared	0.491	0.589	0.573	0.358
Covariates	No	No	No	No

Notes. Estimates are conditional on having obtained a national test result two years before (three for math). In columns (1)–(3), previous test result in reading (grade -2) denotes the grade 2 reading result, the grade 4 reading result, and grade 6 reading result, respectively. In column (4), previous test result in math denotes the grade 3 math score. All specifications include a constant and years fixed effects. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

The analyses above indicate (a strong) persistency in the national test scores of students across grade levels. Figure 2 proceeds to illustrate the distributional transitions of students' reading proficiency between grades 2 and 4 (transitions across other grade levels are similar), i.e. two years apart. As expected, students from the smaller groups in the tails are relatively more mobile, while the greater share of students in the larger groups across the mean remains in the same category in grade four. The far-left bar demonstrates that 30% of the students testing considerably below average (points 1–10, recall that points and norms are based on the 2010 pilot) in the second grade, still test below 10 points in grade four, while 47% advance to the larger below average-group (11–35 points). Four percent test above or considerably above average (66+ points) in grade 4. To the far right, 94% of the students testing considerably above average (91–100 points) in grade two remain above or considerably above average in grade four. Finally, 53% of the students testing on average (36–65 points) in grade two remain in this category in grade four. Slightly more move up the test score distribution (28%) than down (18%). Auxiliary analyses show that children from more favorable socioeconomic backgrounds are more likely to improve from the bottom of the distribution compared to other children. We leave a detailed analysis of the mobility of test scores to future research.

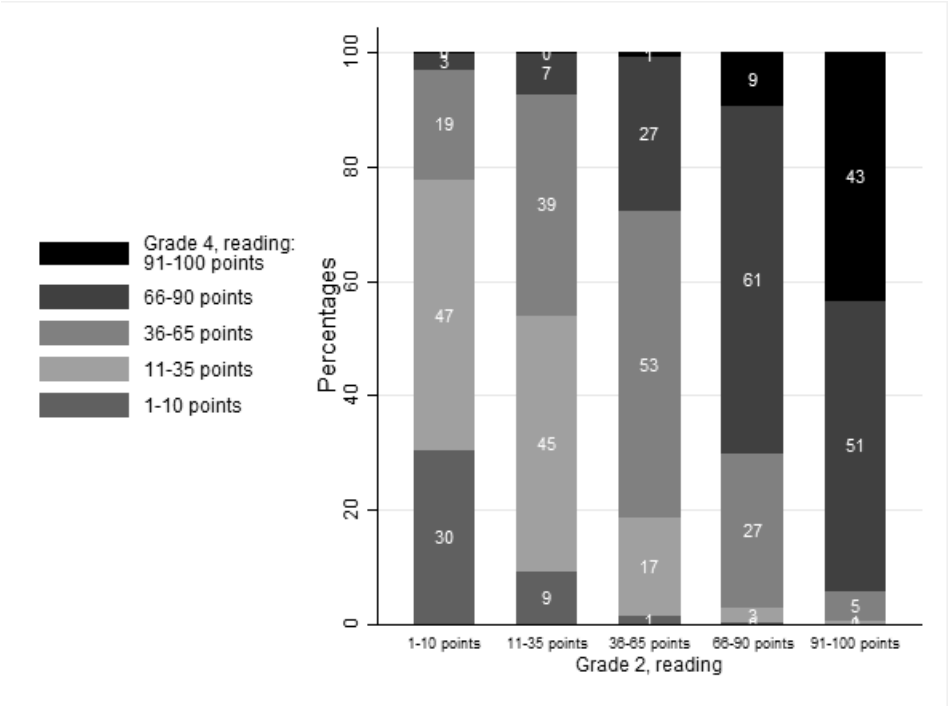

Figure 2. The mobility of students' reading results from grades 2 to 4.
The figure compares student test scores on the second grade reading test to the fourth grade reading scores of the same students (obtained two years later). Points are based on percentiles from the 2010 pilot.

In summary, the results of this section, the predictive ability of the national tests in terms of achievement in exams and the persistency of national test results across grades levels, incite the national tests as policy device for early assessment of students' needs and differentiated instruction. We illustrate this further in the next section.

4.2. Socioeconomic gaps in test scores

A growing body of international literature documents strong associations between students' background characteristics and their performance in school; lower student achievement is associated with low birth weight, being assigned to special needs education and lower socioeconomic status as represented by parents' earnings or education (see e.g. Carneiro and Heckman 2003; Hanushek and Woessmann 2011; Björklund and Salvanes 2011). We find the same patterns in the national test scores, see Table A9 in the Appendix.[21] However, our understanding

21. For all subjects, student and parental background explain 13 to 21% of the variability in the student's national test scores. Similar magnitudes are documented internationally; see e.g.

of how achievement gaps develop though children's schooling is not well documented (see e.g. discussion in Reardon et al. (*forthcoming*)).

In the following, we consider selected background characteristics and illustrate how the average national test score gaps in reading and math develop as students progress through compulsory school. As evident from the figures, one of the main gains from nationwide and repeated testing is the possibility to monitor (mean) achievement of specific groups of students. Hence, potential trajectories may be detected, for example, if the (mean) achievement of different groups of students diverges between grades 4 and 6 it may be suggestive for targeting school resources or intervention policies at this particular grade level.

The left (right) panel of Figure 3 illustrates average student achievement in reading (math) for boys and girls divided by immigration background. In line with international findings, girls generally perform better than boys in reading, while boys outperform girls in math (see e.g. OECD 2015).[22] For students of non-Western background, the gender gap in reading scores converges slightly between grades 2 and 4 and the difference in means disappear in grade 8. Overall, average achievement of students from non-Western countries are considerably lower than others' without any sign of convergence: the reading gap is about three quarters of a standard deviation and in the math gap is about 0.6 SD.

Hanushek and Woessmann (2011) for a review. Comparing Tables A9 and 5, it is worth noting that national test scores, obtained at an earlier grade level, alone explain over 34 percentage points more of the variation in later reading scores compared to all the other student background characteristics combined. However, some of this difference is likely explained by similarity of tasks across tests.

22. Recall that these transfers are relative to other students because of the standardization. Thus, girls may on average have progressed in math proficiency since the previous test but not relatively more than boys have. Furthermore, the graphs will strictly speaking not reveal whether any convergence is caused by the improvement of one group or the deterioration of another. Appendix Figures A1 and A2 replicate the results using test scores in points, which allows one to assess the nature of potential con-/divergences. The patterns are generally similar.

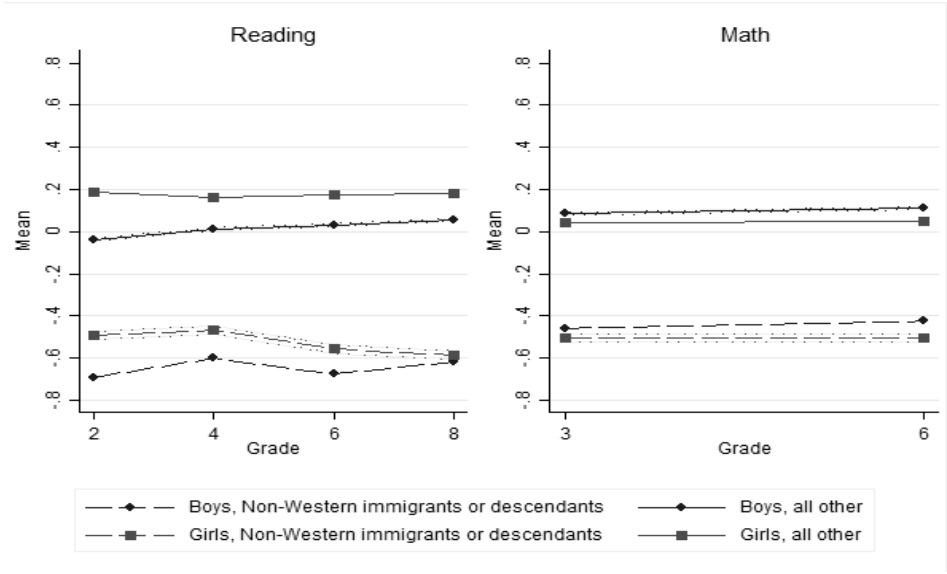

Figure 3. Average national test scores in reading (left panel) and math (right panel) by gender and immigration background

Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010-2013. Approximately 8.6% of the boys and 8.8% of the girls are Non-Western immigrants or descendants.

If we consider students by their socioeconomic status as proxied by parental education and income, there is a clear and stable social gradient in average test scores, see Figures 4 and 5, respectively. This pattern is observed for reading (left panels) as well as for math scores (right panels). In standard deviation metrics, children from families where the mother has obtained no more than compulsory schooling perform on average one quarter of a SD below children from families where the mother has vocational training, and more than one SD below families where the mother holds a higher tertiary degree.

Figure 5 illustrates a one SD difference in mean test scores between students from the bottom and top income quartile. The magnitudes of these socioeconomic achievement gaps are comparable to international findings. For example, Carneiro and Heckman (2003) document similar stable test score gaps from the age 6 to 12 years by family income quartiles and race. Reardon et al. (2011) compare income achievement gaps across different longitudinal studies in the US: the differences in standardized test scores between the 90th and 10th income percentile families are all above 1 SD.

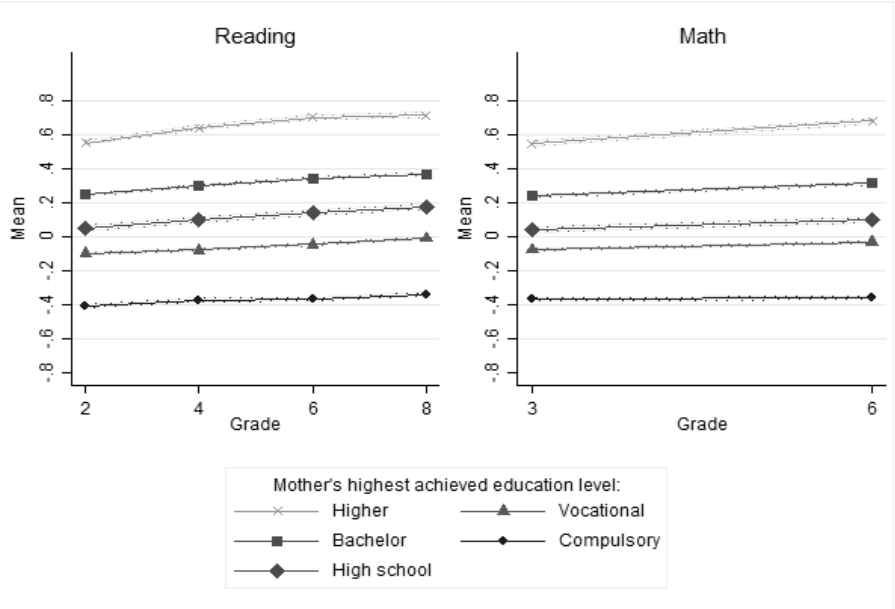

Figure 4. Average national test scores in reading (left) and math (right panel) by mother's educational attainment

Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010-2013. Approximately 23% of the sample have compulsory education, 8% have high school, 36% have vocational education or training, 26% have a bachelor degree, and 8% have completed a higher education.

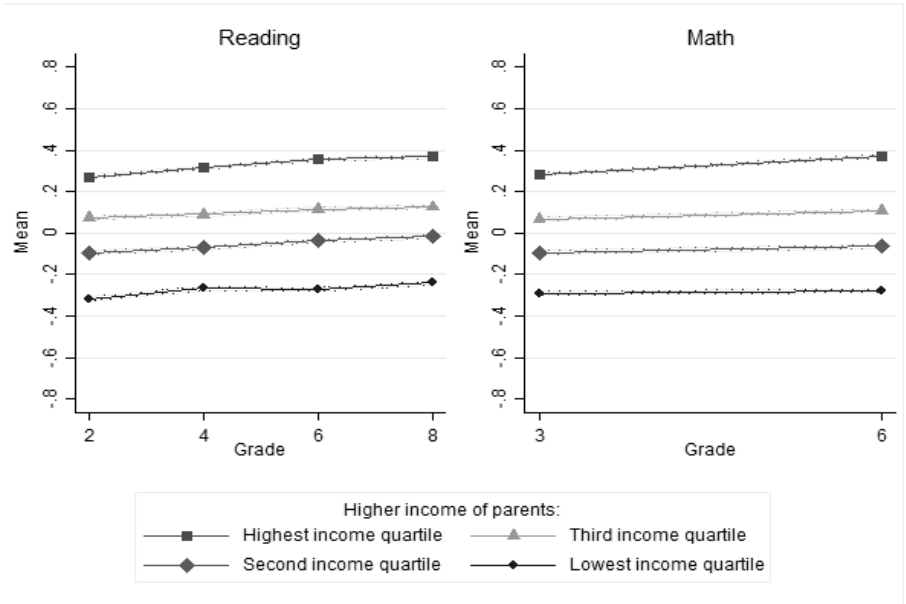

Figure 5. Average national test scores in reading (left panel) and math (right panel) by parental earnings

Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010-2013. Approximately one quarter of the sample are in each of the income groups.

Lastly, we segregate students by their special education needs. As these are determined in collaboration with the school, they are directly observable to schools. Progressions such as these could serve as a benchmark for schools evaluating their programs for inclusive education. In Figure 6, the blue line denotes average student achievement for students with no documented special education needs at age 8. All referrals are measured at age 8, i.e. before the earliest national test in grade 2.[23] In grade 2, the average test score gap is smallest between non-referred students and students referred with physical disabilities. The gap is slightly larger for students with mental disabilities (0.4 SD) while it is roughly the same for students with social and other/unspecified disabilities (both around 0.6 SD). The largest gap is for students with learning disabilities; they score nearly one SD below their peers with no special education needs at age 8. This is comparable to Floridian data showing that students taught in a mix of mainstream and special education classes (comparable to our sample students with special education

23. At age 8, 5.0% of the 2nd grade sample, 4.8% of the 4th grade sample, and 3.7% of the 6th grade sample have a referral for special needs education. Among these, the primary causes are distributed as follows: 52% have learning disabilities, 40% other/unspecified disabilities, 4% mental disabilities, 2% social disabilities, and 2% physical disabilities.

needs) on average score 0.9 SD below their peers in math and reading (Feng and Sass 2013). Overall, the test score gaps in Figure 6 decrease by 0.1-0.2 SD across grades. However, (part of) this convergence is likely driven by students with the most detrimental disabilities transferring to segregated special education classes or schools across grades, thus, dropping out of our sample. Figure 6 suggests that students with documented disabilities at age 8, despite being assigned to special needs education, only to a small degree catch up with their average peers during schooling. More importantly, though, the test score gaps do not widen. This is in contrast to American literature. For example, Hanushek et al. (2002) show that the gap in math between regular and emotionally disabled students widens from 0.69 SD in grade 4 to 0.95 SD in grade 7, and the test score gap between regular and learning disabled students widens from 0.84 SD in grade 4 to 1.07 SD in grade 7 (similar for reading).

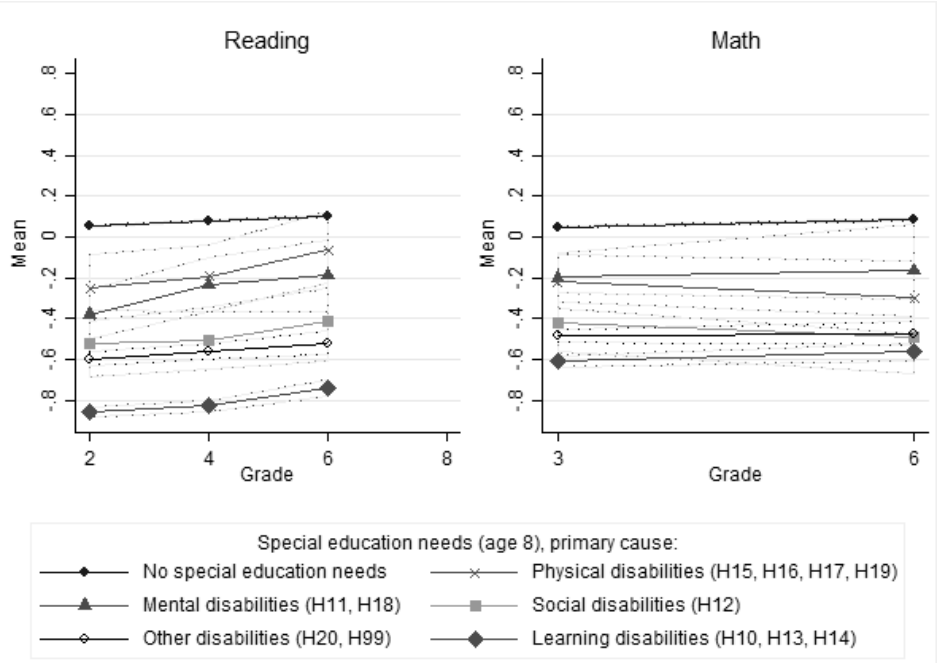

Figure 6. Average national test scores in reading (left panel) and math (right panel) by special education needs
Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010–2013. Grade 8 observations are omitted due to lack of special education information at age 8 for this age group. Disability categories are aggregates of the 12 official ICD-10 categories in parentheses.

In conclusion, the findings of this paper suggest that inequalities in student test scores are present—and identifiable—from the first test in the second grade, per-

sist throughout compulsory schooling, and are associated with final exam results. However, this also points toward exploiting the national test results to evaluate and compare the effect of different school interventions: benchmarking against policy-relevant achievement gaps and observed effect sizes for different interventions may support and improve policy decisions. For example, Andersen et al. (2016b) show that a teaching assistant in sixth grade classrooms can close one third of the achievement gap between students of low and high-educated parents documented in this paper. On the other hand, Nandrup (2016) finds that class sizes do not generally affect the achievement gaps in reading and math scores, although, a one-student decrease in class size on average increases test scores by around 0.01 SD. In addition, interventions targeted toward specific groups of students may in fact widen the gap, if the intervention is as least as beneficial for the class peers. Evidence from Andersen et al. (2016a) suggest that increasing weekly instruction time in Danish may widen the reading gap between students of Danish and non-Western origin by as much as 25%.

5. Concluding remarks

This paper describes the format of the Danish national tests and provides empirical analyses of the relationship between student national test scores, student background characteristics, ninth grade exam results, and enrollment in further education. We focus on two key issues: the predictive validity of the national test results and the knowledge gained from national and mandatory testing.

First, we consider the persistence and predictive validity of the national test results in terms of test scores and attainment at later stages of education. We utilize the complete set of student test scores from the first four years of the program to obtain high statistical power. We find that the national test scores in early grades are strong predictors for students' test scores in later grades. We further demonstrate significant empirical associations between reading scores and later, more common measures of success, such as the ninth grade exam results and enrollment in upper secondary education (general or vocational). Overall, we conclude that the national tests are able to measure skills that are highly correlated with the skills measured by the ninth grade examination. A next important step is to validate whether improving students' national test scores, e.g. by and intervention in the fourth grade, also translates into improved ninth grade exam results (or other educational outcomes measures).

Second, we provide examples of the knowledge gained by yearly, mandatory, and standardized testing of students. We document a socioeconomic gradient in student test scores comparable to the international literature and graphically illustrate considerable, stable test score gaps for children with special education needs and poorer socioeconomic backgrounds across all grade levels. Although, poten-

tially coming with the costs of added stressors for students, instruction time (and preparation time for teachers) spent testing instead of learning, and potentially demotivating low-achieving students, these insights are valuable from a policy perspective for addressing social mobility at different stages of education.

One of the national aims of the Danish public school system is to reduce the socioeconomic gradient in student achievement. Ideally then, any potential socioeconomic achievement gap existing at school entry will diminish with schooling. Our results suggest that this may not be the case for Denmark. Of course, we do not know the counterfactual—what would have happened in a world without the current level (and adjustment) of school inputs. Would the current gaps in achievement be even larger, and growing? However, with the introduction of standardized and objective national tests, at multiple grade levels, it is now possible to evaluate the impact of school interventions and programs aimed at improving achievement for specific student groups or grade levels (see e.g. Andersen et al. 2016a; Andersen et al. 2016b; Nandrup 2016). Thereby improving the possibility for developing evidence-based policies to narrow achievement gaps in Danish public schools.

References

Andersen, S. C., Humlum, M., and Nandrup, A. B. (2016a), »Increasing instruction time in school does increase learning«, *Proceedings of the National Academy of Sciences of The United States of America*, 113(27): 7481-7484.

Andersen, S. C., Beuchert, L.V., Nielsen, H. S., and Thomsen, M. K. (2016b), »The effect of teacher's aides in the classroom: Evidence from a randomized trial.« SSRN: https://ssrn.com/abstract=2626677

Andersen, S. C. and Nielsen, H. S. (2016), »The Positive Effects of Nationwide Testing on Student Achievement in a Low-Stakes System«. SSRN: https://ssrn.com/abstract =2628809

Björklund, A. and Salvanes, K. G. (2011), »Education and Family Background: Mechanisms and Policies«, *Handbook of the Economics of Education*, 3.

Bond, T. and Fox, C. (2007), »Applying the Rasch Model«, Second edition. *Routledge*.

Carneiro, P. and Heckman, J. (2003), »Human Capital Policy«, in J. Heckman and A. Krueger (eds.), *Inequality in America: What Role for Human Capital Policies?* (Cambridge, MA: MIT Press, 2003).

Dee, T. (2007), »Teachers and the Gender Gaps in Student Achievement«, *The Journal of Human Resources,* 42(3): 528-554.

Downey, D. and Pribesh, S. (2004), »When Race Matters: Teachers' Evaluations of Students' Classroom Behavior«, *Sociology of Education,* **77**: 267-282.

Feng, L. and Sass, T. R. (2013), »What makes special-education teachers special? Teacher training and achievement of students with disabilities«, *Economics of Education Review,* 36: 122-134.

Figlio, D. and Loeb, S. (2011), »School Accountability« Ch. 8 in E. A. Hanushek, S. Machin and L. Woessmann (eds.), *Handbooks in Economics, 3*.

Hanushek, E.A., J.F. Kain, and Rivkin, S. G. (2002), »Inferring Program Effects for Special Populations: Does Special Education Raise Achievement for Students with Disabilities?«, *The Review of Economics and Statistics,* 84(4): 584-599.

Hanushek, E.A and Woessmann, L. (2011), »The Economics of International Differences in Education Achievement«, *Handbook of the Economics of Education,* 3.

Hinnerich, B. T. and Vlachos, J. (2013), »Systematiska skillnader mellan interna och externa bedömningar av nationella prov – en uppföljningsrapport«, Report to the Swedish School Inspectorate, 2013.

Hinnerich, B. T. and Vlachos, J. (2017), »The Impact of Upper-Secondary Voucher School Attendance on Student Achievement. Swedish Evidence Using External and Internal Evaluations«, IFN Working Paper No. 1127. SSRN: https://ssrn.com/abstract=2943291

Humlum, M.K. og T.P. Jensen (2010), »Frafald på de erhvervsfaglige uddannelser - Hvad karakteriserer de frafaldstruede unge?« *AKF Working paper*.

Hvidtfeldt, C. and T. Tranæs (2013), »Folkeskolekarakterer og Succes på Erhvervsuddannelserne«, *Rockwool Fondens Forskningsenhed*, Working paper no. 29.

Jacob, B., and Rothstein, J. (2016), »The Measurement of Student Ability in Modern Assessment Systems«, NBER Working Paper No. 22434. Published in: *The Journal of Economic Perspectives*, 30(3): 85-107.

Nandrup, A.B. (2016), »Do Class Size Effects Differ Across Grades?«, *Education Economics*, 24(1): 83-95.

OECD (2004), »Reviews of National Policies for Education: Denmark – Lessons from PISA 2000«, *OECD Publishing,* Paris. DOI: http://dx.doi.org/10.1787/9789264017948-en.

OECD (2013), »Synergies for Better Learning: An International Perspective on Evaluation and Assessment«, *OECD Publishing*, Paris. DOI: http://dx.doi.org/10.1787/9789264190658-en

OECD (2015), »The ABC of gender equality in education: Aptitude, behavior, confidence«, Pisa, OECD Publishing. DOI: http://dx.doi.org/10.1787/9789264229945-en (March 5, 2015).

The Public School Act (2013), Folkeskoleloven, Available at: https://www.retsinformation.dk/ Forms/r0710.aspx?id=145631

Rambøll (2013), »Evalueringen af de nationale test i Folkeskolen«. Available at: http://uvm.dk/~/UVM-DK/Content/News/Udd/Folke/2013/Okt/~/media/UVM/Filer/Udd/Paa%20tvaers/Priser/131010%20Evaluering%20af%20de%20nationale%20test_rapport.ashx

Rangvid, B. S. (2015), »Systematic differences across evaluation schemes and educational choice«, *Economics of Education Review*, 48: 41-55.

Reardon, S.F. (2011), »The widening academic achievement gap between the rich and the poor: New evidence and possible explanations.« In R. Murnane & G. Duncan (Eds.), Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children. New York: Russell Sage Foundation Press

Reardon, S.F., Robinson, J.P., and Weathers, E.S. (Forthcoming), »Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps.« In H. A. Ladd & E. B. Fiske (Eds.), Handbook of Research in Education Finance and Policy (Second ed.). Lawrence Erlbaum.

Rasch, G. (1960), »Probabilistic models for some intelligence and attainment tests«, *Copenhagen: Danske Paedagogiske Institut*.

Review (2007) (Sørensen, H., Norrild, P., Petersen, D. K., Elbro, C., Mogensen, A., Hansen, K.F., et al.), »De nationale IT-baserede test i folkeskolen – rapport fra REVIEW-panelet«, *Undervisningsministeriet*, Devoteam Consulting A/S. Available at: http://www.folkeskolen.dk/ ~/1/7/63134-v7-reviewafdenationaleit-baseredetest.pdf (March 2012)

Undervisningsministeriet (2012), »Den Adaptive Algoritme i De Nationale Test«, (L. Strange, C. Jensen, H. Albeck and J. Lund). *Undervisningsministeriet, Statistik*

og Analyse (UNI-C). Available at:http://uvm.dk/~/media/UVM/Filer/Udd/Folke/PDF14/Jan/140127%20Notat%20om%20den%20adaptive%20algoritme%20i%20de%20nationale%20test.pdf (April 2015)

Undervisningsministeriet (2014), »Statistisk usikkerhed i de nationale test«, *Undervisningsministeriet, Statistik og Analyse (UNI-C)*. Available at: http://uvm.dk/Aktuelt/~/UVM-DK/Content/News/Udd/Folke/2014/Jan/140127-Statistisk-sikkerhed-i-de-nationale-test.

Undervisningsministeriet (2015), »Den adaptive algoritme i De Nationale Test«, *Undervisningsministeriet, Styrelsen for IT og læring (STIL)*. Available at: http://uvm.dk/~/media/UVM/Filer/Udd/Folke/PDF15/Jan/150128%20Den%20adaptive%20algoritme%20i%20De%20Nationale%20Test.pdf (June 2015)

Wandall, J. (2011), »National Test in Denmark – CAT as a Pedagogic Tool«, *Journal of applied testing and technology,* 12(3).

Appendix

Table A1. Subject-specific cognitive domains (profile areas) of the national tests

Subject	Profile Area 1 (domain 1)	Profile Area 2 (domain 2)	Profile Area 3 (domain 3)
Danish, reading	Language comprehension (Sprogforståelse)	Decoding (Afkodning)	Reading comprehension (Tekstforståelse)
Mathematics	Numbers and algebra (Tal og algebra)	Geometry (Geometri)	Mathematics in use (Matematik i anvendelse)
Physics/Chemistry	Energy (Energi og energiomsætning)	Phenomena, substances and materials (Fænomener, stoffer og materialer)	Applications and perspectives (Anvendelser og perspektiver)
English	Reading (Læsning)	Vocabulary (Ordforråd)	Language and linguistic usages (Sprog og sprogbrug)
Geography	Natural geography (Naturgrundlaget)	Cultural geography (Kulturgeografi)	Applied geography (At bruge geografien)
Biology	The living organism (Den levende organisme)	The interplay of living organisms (Levende organismers samspil med hinanden og deres omgivelser)	Applied biology (At bruge biologien: Biologiens anvendelse, tankegange og arbejdsmetoder)
Danish as second language	Vocabulary (Ordforråd)	Language and linguistic usages (Sprog og sprogbrug)	Reading comprehension (Læseforståelse)

Notes. English translation provided by Wandall (2011), however Danish as second language is partly translated by the authors. Examples of items are available at https://demo.testogprøver.dk/.

Table A2. Sample means of background characteristics

	Full sample	
	Mean	Std. Dev.
Student characteristics		
Girl (0/1)	0.495	
Western immigrant/descendant (0/1)	0.010	
Non-Western immigrant/descendant (0/1)	0.090	
Low birthweight (<2500 g)	0.106	
First-born (0/1)	0.431	
Second-born (0/1)	0.372	
Third-born or later (0/1)	0.187	
Multiple borns (e.g. twins) (0/1)	0.038	
Born in the first quarter (0/1)	0.244	
- in the second quater	0.253	
- in the third quater	0.263	
- in the fourth quater	0.236	
Age indicators (omitted here)		
Psychiatric diagnosed (0/1)	0.018	
ADHD diagnosed, age 8 (0/1)	0.003	
No. of school transfer within the last 2 years	0.191	
Referred to special needs education in the previous year (0/1)		
- Learning disability	0.032	
- Mental disability	0.002	
- Social disability	0.001	
- Physical disability	0.001	
- Other	0.023	
Familiy information (year 5)		
No. of siblings	1.218	0.866
Single mom (0/1)	0.148	
Mother's logearnings	9.634	4.854
Mother has negative earnings	0.166	
Mother's age (years)	32.639	7.544
Mother's education: None or missing (0/1)	0.051	
Mother's education: Compulsory or high school (0/1)	0.271	
Mother's education: Vocational (0/1)	0.361	
Mother's education: Bachelor (0/1)	0.246	
Mother's education: Higher (0/1)	0.071	
Father's logearnings	10.275	4.800
Father has negative earnings (0/1)	0.129	
Father's age (years)	34.693	9.406
Father's education: None or missing (0/1)	0.066	
Father's education: Compulsory or high school (0/1)	0.248	
Father's education: Vocational (0/1)	0.412	
Father's education: Bachelor (0/1)	0.181	
Father's education: Higher (0/1)	0.093	

School information		
School size (#students)	465.461	177.810
Class size (#students)	21.729	3.953
Capital area school (0/1)	0.064	
Bigger city school (0/1)	0.115	
N	2,116,150	

Notes. All student and family characteristics are measured at age 5 unless otherwise indicated. When covariates are included in regressions, relevant indicators for missing covariates are always included. We set logearnings equal 0 for parents earning zero earnings.

Table A3. Correlations between each of the three cognitive domains by subject (raw logit scores)

Cognitive domains:	Domain 1	Domain 2	Domain 3	Domain 1	Domain 2	Domain 3	Domain 1	Domain 2	Domain 3
	Reading, grade 2			Reading, grade 4			Math, grade 3		
Domain 1	1.00	0.57	0.59	1.00	0.64	0.69	1.00	0.65	0.73
Domain 2	-	1.00	0.79	-	1.00	0.68	-	1.00	0.66
Domain 3	-	-	1.00	-	-	1.00	-	-	1.00
	Reading, grade 6			Reading, grade 8			Math, grade 6		
Domain 1	1.00	0.61	0.63	1.00	0.55	0.61	1.00	0.58	0.66
Domain 2	-	1.00	0.69	-	1.00	0.62	-	1.00	0.60
Domain 3	-	-	1.00	-	-	1.00	-	-	1.00

Notes. For each reading and math test, the table shows the raw correlations between the test scores of each cognitive domain (i.e. profile area 1, 2, and 3, see Appendix Table A1). The sample includes all national test results for public school students from 2010–2013.

Table A4. Correlations between the constructed standardized average and each cognitive domain by subject (standardized logit scores).

Cognitive domains:	Domain 1	Domain 2	Domain 3
Standardized average:			
Reading, grade 2	0.83	0.90	0.90
Reading, grade 4	0.88	0.87	0.90
Reading, grade 6	0.85	0.88	0.89
Reading, grade 8	0.83	0.84	0.87
Math, grade 3	0.89	0.87	0.90
Math, grade 6	0.87	0.84	0.87

Notes. For each reading and math test, the table shows the raw correlations between the test scores of each cognitive domain (i.e. profile area 1, 2, and 3, see Appendix Table A1) and the standardized average test score. Standardized averages are calculated as described in section 3.2. The sample includes all national test results for public school students from 2010–2013.

Table A5. The number and percentages of students enrolled in public school with a national test result by subject and year

	2010		2011		2012		2013		Total
	Test obs.	%	Test obs.	%	Test obs.	%	Test obs.	%	Test obs.
Reading, grade 2	46,578	85.6	50,777	96.8	51,873	97.8	50,763	96.7	199,991
Reading, grade 4	47,528	86.3	53,272	96.9	52,009	97.6	50,014	96.9	202,823
Reading, grade 6	48,729	87.3	52,530	97.1	52,131	97.5	51,504	96.8	204,894
Reading, grade 8	44,098	81.6	48,339	93.9	49,173	95.2	47,209	94.2	188,819
Math, grade 3	48,999	87.9	52,102	96.0	50,939	97.7	51,538	96.4	203,578
Math, grade 6	48,923	87.7	52,368	96.8	52,128	97.5	51,391	96.6	204,810
English, grade 7	45,405	84.1	51,035	94.5	50,067	95.7	48,678	94.2	195,185
Physics/Chemistry, grade 8	44,504	82.4	47,715	92.6	48,442	93.8	45,970	91.7	186,631
Biology, grade 8	43,819	81.1	47,711	92.6	48,385	93.7	46,255	92.3	186,170
Geography, grade 8	44,180	81.8	47,788	92.8	48,323	93.6	46,055	91.9	186,346

Notes. Test obs. denotes the total number of students tested in a given grade and year. The corresponding percentages indicate the share of tested students to the total number of students enrolled in public school in a given grade and year.

Table A6. OLS estimates: ninth grade exam marks on national test scores, with baseline covariates

	(1) GPA, Danish	(2) GPA Danish	(3) GPA math	(4) Exit exam, English	(5) Exit exam, physics
National test results:					
Reading, grade 6	1.7463 ***				
	(0.014)				
Reading, grade 8		1.762 ***			
		(0.010)			
Math, grade 6			1.950 ***		
			(0.018)		
English, grade 7				2.251 ***	
				(0.016)	
Physics/chemistry, grade 8					1.645 ***
					(0.022)
Selected covariates					
Girl	1.083 ***	1.036 ***	-0.282 ***	0.419 ***	0.676 ***
	(0.018)	(0.011)	(0.021)	(0.021)	(0.021)
Western immigrant/ descendant	0.148	-0.065	0.011	-0.071	-0.198 **
	(0.094)	(0.050)	(0.117)	(0.101)	(0.099)
Non-Western immigrant/descendant	0.335 ***	0.124 ***	-0.299 ***	0.249 ***	-0.080
	(0.047)	(0.028)	(0.050)	(0.051)	(0.052)
Low birthweight (<2500)	-0.094 **	0.094 ***	-0.149 ***	0.059	-0.124 **
	(0.043)	(0.024)	(0.052)	(0.050)	(0.048)
No. of siblings	0.068 ***	0.043 ***	0.103 ***	0.028 *	0.103 ***
	(0.013)	(0.008)	(0.015)	(0.015)	(0.014)
ADHD diagnosed	-0.711 ***	-0.240 **	-0.118	-0.319	-0.391 *
	(0.230)	(0.120)	(0242)	(0232)	(0233)
Special education need, primary cause					
Learning disability	-0.631 ***	-0.635 ***	-0.868 ***	-0.629 ***	-0.828 ***
	(0.063)	(0.034)	(0.070)	(0.061)	(0.059)
Mental disability	-0.014	-0.373 ***	0.306	-0.266	-0.247
	(0.218)	(0.112)	(0.358)	(0.207)	(0.206)
Social disability	-0.593 **	-0.405 **	-0.659	-0.410	-0.910 **
	(0.285)	(0.197)	(0.429)	(0.406)	(0.388)
Physical disability	-0.337	0.151	-0.328	-0.302	0.237
	(0.270)	(0.214)	(0.274)	(0.441)	(0.377)
Other	-0.575 ***	-0.522 ***	-0.628 ***	-0.335 ***	-0.795 ***
	(0.072)	(0.040)	(0.084)	(0.078)	(0.072)
Family information					
Single mother	-0.219 ***	-0.180 ***	-0.322 ***	-0.074 **	-0.361 ***
	(0.025)	(0.015)	(0.031)	(0.029)	(0.028)
Mother's age	0.024 ***	0.021 ***	0.022 ***	0.032 ***	0.025 ***
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
Father's education					
≤ High school	0.028	-0.011	-0.022	-0.086	0.023

	(0.071)	(0.033)	(0.074)	(0.074)	(0.069)
Vocational	0.194 ***	0.086 **	0.197 ***	0.079	0.213 ***
	(0.071)	(0.034)	(0.075)	(0.074)	(0.070)
Bachelor	0.490 ***	0.381 ***	0.542 ***	0.409 ***	0.700 ***
	(0.072)	(0.035)	(0.077)	(0.075)	(0.072)
Higher	0.673 ***	0.516 ***	0.764 ***	0.446 ***	0.898 ***
	(0.076)	(0.037)	(0.082)	(0.079)	(0.077)
Father's logearnings	0.062 ***	0.060 ***	0.044 ***	0.035 ***	0.074 ***
	(0.011)	(0.007)	(0.014)	(0.013)	(0.012)
Capital area school	-0.118 *	-0.088 **	-0.018	-0.192 **	0.205 **
	(0.060)	(0.039)	(0.069)	(0.077)	(0.081)
Observations	46,728	138,970	46,484	91,331	134,929
R-squared	0.582	0.581	0.533	0.445	0.275

Notes. Selected coefficients are shown. In addition to the control variables listed in the table, all specifications include a constant, year fixed effect, and the remaining controls from Table A2. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table A7. OLS estimates: enrollment and completion of upper secondary education (general or vocational) on national test result, with baseline covariates

	(1)	(2)	(3)
	Enrolled	Completed or enrolled	Completed general upper secondary
	2 years after compulsory school	4 years after compulsory school	4 years after compulsory school
National test result:			
Reading, grade 8	0.161***	0.086***	0.157***
	(0.002)	(0.002)	(0.003)
Observations	141,558	44,073	44,073
Cohorts	3	1	1
Mean outcome	0.776	0.830	0.361
R-squared	0.228	0.119	0.141
Covariates	Yes	Yes	Yes

Notes. All specifications include a constant, year fixed effect, and all controls from Table A2. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table A8. OLS estimates: National test results explained by previous test result in the same subject, with baseline covariates

	(1) Reading, grade 4	(2) Reading, grade 6	(3) Reading, grade 8	(4) Math, grade 6
Previous national test results:				
Reading, (grade -2)	0.621***	0.703***	0.690***	
	(0.004)	(0.004)	(0.004)	
Math, grade 3				0.517***
				(0.006)
Observations	90.194	92.922	87.110	43.827
Cohorts	2	2	2	1
Mean outcome	0.058	0.067	0.089	0.067
R-squared	0.516	0.605	0.589	0.403
Covariates	Yes	Yes	Yes	Yes

Notes. Estimates are conditional on having obtained a national test result two years before (three for math). In columns (1)–(3), previous test result in reading, (grade -2) denotes the grade 2 reading result, the grade 4 reading result, and grade 6 reading result, respectively. In column (4), previous test result in math denotes the grade 3 math score. All specifications include a constant, years fixed effects, and all controls from Table A2. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table A9. OLS estimates: national test results explained by parental and student characteristics, with baseline covariates - linguistic tests

Select covariates	(1) Reading, grade 2	(2) Reading, grade 4	(3) Reading, grade 6	(4) Reading, grade 8	(5) Reading, grade 7
Girl	0.215 ***	0.117 ***	0.108 ***	0.081 ***	0.078 ***
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)
Western immigrant/descendant	-0.127 ***	-0.185 ***	-0.256 ***	-0.268 ***	-0.064 **
	(0.022)	(0.022)	(0.024)	(0.025)	(0.027)
Non-Western immigrant/descendant	-0.372 ***	-0.349 ***	-0.444 ***	-0.445 ***	-0.136 ***
	(0.015)	(0.014)	(0.015)	(0.015)	(0.015)
Low birthweight (<2500)	-0.107 ***	-0.060 ***	-0.043 ***	-0.043 ***	-0.047 ***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
No. of siblings	-0.007 *	-0.006 *	0.001	-0.000	-0.033 ***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
ADHD diagnosed	-0.082 **	0.034	0.037	0.038	0.065
	(0.042)	(0.042)	(0.045)	(0.053)	(0.052)
Special education needs, primary cause					
Learning disability	-0.529 ***	-0.867 ***	-0.917 ***	-1.004 ***	-0.916 ***
	(0.029)	(0.016)	(0.014)	(0.019)	(0.013)
Mental disability	-0.134 **	-0.395 ***	-0.332 ***	-0.301 ***	-0.199 ***
	(0.065)	(0.060)	(0.051)	(0.056)	(0.049)

Social disability	-0.249 ***	-0.364 ***	-0.492 ***	-0.640 ***	-0.370 ***
	(0.076)	(0.089)	(0.071)	(0.117)	(0.076)
Physical disability	-0.251 ***	-0.273 ***	-0.339 ***	-0.439 ***	-0.379 ***
	(0.071)	(0.098)	(0.081)	(0.099)	(0.074)
Other	-0.362 ***	-0.581 ***	-0.609 ***	-0.635 ***	-0.577 ***
	(0.028)	(0(0.017) .024)	(0.020)	(0.026)	(0.018)
Family informations					
Single mom	-0.067 ***	-0.028 ***	-0.023 ***	-0.018 ***	-0.010
	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)
Mother's age	0.011 ***	0.014 ***	0.015 ***	0.015 ***	0.016 ***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Father's education					
≤ High school	0.002	-0.019	0.008	0.038 **	-0.008
	(0.017)	(0.016)	(0.017)	(0.018)	(0.017)
Vocational	0.044 **	0.013	0.048 ***	0.070 ***	-0.006
	(0.018)	(0.016)	(0.017)	(0.018)	(0.018)
Bachelor	0.245 ***	0.219 ***	0.257 ***	0.270 ***	0.238 ***
	(0.018)	(0.017)	(0.018)	(0.019)	(0.018)
Higher	0.359 ***	0.348 ***	0.388 ***	0.395 ***	0.382 ***
	(0.019)	(0.018)	(0.019)	(0.020)	(0.020)
Fathers logearnings	0.025 ***	0.022 ***	0.018 ***	0.019 ***	0.028 ***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Capital area school	-0.88 ***	-0.021	-0,008	-0.058 **	-0.132 ***
	(0.024)	(0.028)	(0.029)	(0.028)	(0.026)
Observations	199,991	202,823	204,894	188,819	195,185
Mean outcome	0.017	0.032	0.039	0.052	0.025
R-suared	0.148	0.181	0.206	2.200	0.168

Notes, see next page (continued)

Table A9 (cont.). OLS estimates: national test results explained by parental and student characteristics, with baseline covariates - math and science tests

Select covariates	(6) Math, grade 3	(7) Math, grade 6	(8) Physics/ chemistry, grade 8	(9) Biology, grade 8	(10) Geography, grade 8
Girl	-0.064 ***	-0.090 ***	-0.254 ***	-0.034 ***	-0.188 ***
	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)
Western immigrant/ descendant	0.028	-0.047 **	-0.040	-0.129 ***	-0.117 ***
	(0.023)	(0.023)	(0.026)	(0.025)	(0.024)
Non-Western immi-grant/descendant	-0.250 ***	´0.225 ***	-0.269 ***	-0.426 ***	-0.299 ***
	(0.015)	(0.014)	(0.014)	(0.013)	(0.014)
Low birthweight (<2500)	-0.124 ***	-0.116 ***	-0.036 ***	-0.023 **	-0.064 ***
	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)
No. of siblings	0.018 ***	0.032 ***	0.030 ***	0.026 ***	0.025 ***
	(0.004)	(0.004)	(0.003)	(0.004)	(0.004)
ADHD diagnosed	-0.056	-0.121 **	0.025	0.031	0.019
	(0.041)	(0.051)	(0.055)	(0.071)	(0.057)

Special education needs, primary cause					
Learning disability	-0.545 ***	-0.590 ***	-0.440 ***	-0.551 ***	-0.609 ***
	(0.018)	(0.012)	(0.014)	(0.015)	(0.015)
Mental disability	-0.284 ***	-0.301 ***	-0.226 ***	-0.144 **	-0.223 ***
	(0.057)	(0.060)	(0.070)	(0.068)	(0.058)
Social disability	-0.286 ***	-0.521 ***	-0.440 ***	-0.636 ***	-0.669 ***
	(0.084)	(0.074)	(0.077)	(0.151)	(0.203)
Physical disability	-0.152 *	-0.187 **	-0.167	-0.298 ***	-0.275 **
	(0.080)	(0.090)	(0.115)	(0.108)	(0.111)
Other	-0.413 ***	-0.491 ***	-0.341 ***	-0.421 ***	-0.444 ***
	(0.027)	(0.019)	(0.022)	(0.023)	(0.023)
Family information					
Single mom	-0.077 ***	-0.102 ***	-0.097 ***	-0.067 ***	-0.099 ***
	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
Mother's age	0.009 ***	0.011 ***	0.013 ***	0.016 ***	0.018 ***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Father's education					
≤ High school	-0.026	-0.002	0.006	0.025	0.014
	(0.018)	(0.017)	(0.018)	(0.017)	(0.018)
Vocational	0.026	0.071 ***	0.040 **	0.063 ***	0.065 ***
	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)
Bachelor	0.213 ***	0.269 ***	0.245 ***	0.278 ***	0.282
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
Higher	0.339 ***	0.421 ***	0.403 ***	0.431 ***	0.436 ***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Father's logearnings	0.036 ***	0.034 ***	0.013 ***	0.013 ***	0.021 ***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Capital area school	-0.184 ***	-0.136 ***	-0.134 ***	0.037	-0.139 ***
	(0.030)	(0.030)	(0.029)	(0.028)	(0.026)
Observations	203.578	204,810	186,631	186,170	186,346
Mean outcome	0.017	0.034	0.020	0.024	0.026
R-squared	0.125	0.162	0.135	0.154	0.172

Notes. Selected covariates are shown. In addition to the control variables listed in the table, all specification include a constant, year fixed effect, and the remaining controls from Table A2. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

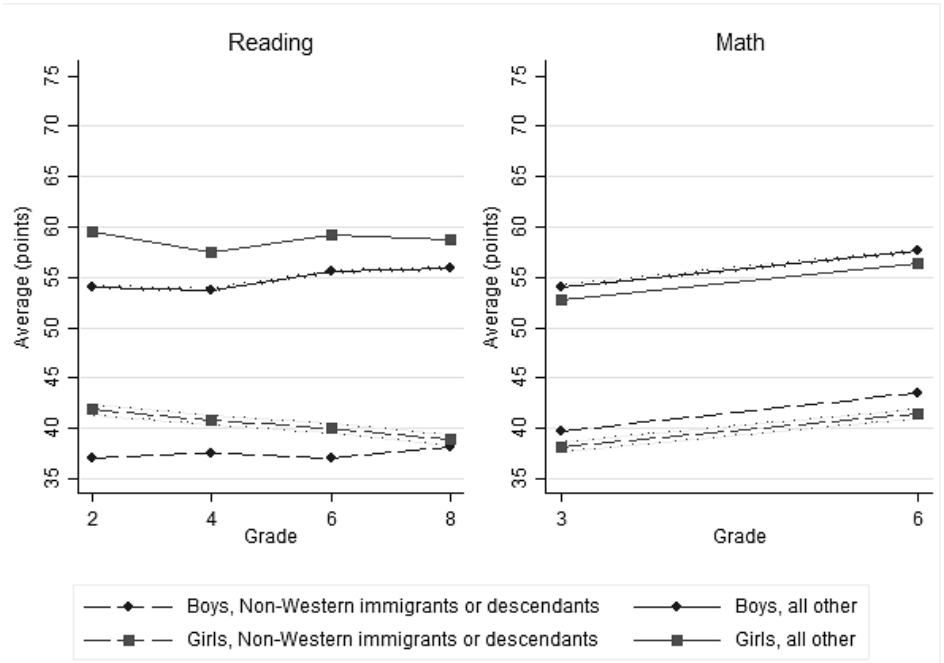

Figure A1. Average national test score (points) in reading (left panel) and math (right panel) by gender and immigration background
Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010–2013. Approximately 8.6% of the boys and 8.8% of the girls are Non-Western immigrants or descendants.

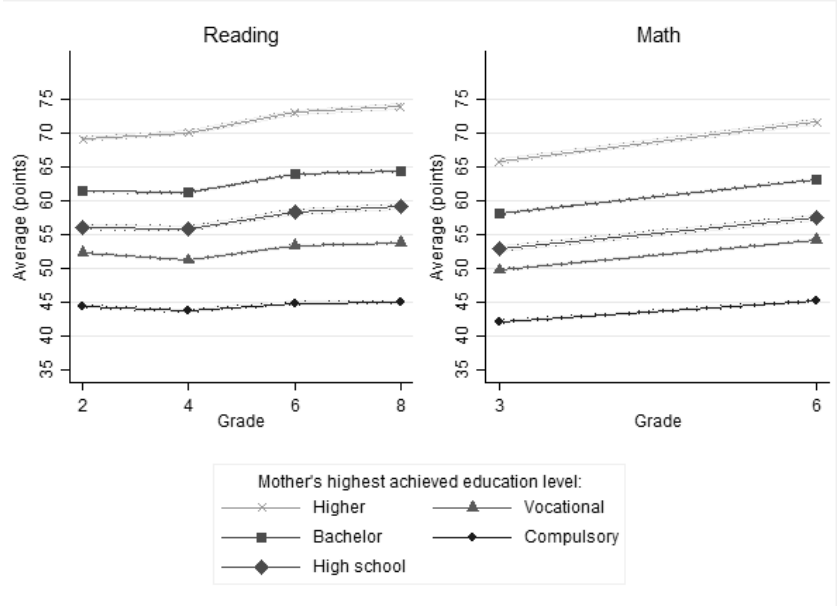

Figure A2. Average national test score (points) in reading (left panel) and math (right panel) by mother's educational attainment

Dotted lines indicate the 95% confidence bands. The figure includes all test results for public school students from the national tests in 2010–2013. Approximately 23% of the sample have compulsory education, 8% have high school, 36% have vocational education or training, 26% have a bachelor degree, and 8% have completed a higher education.