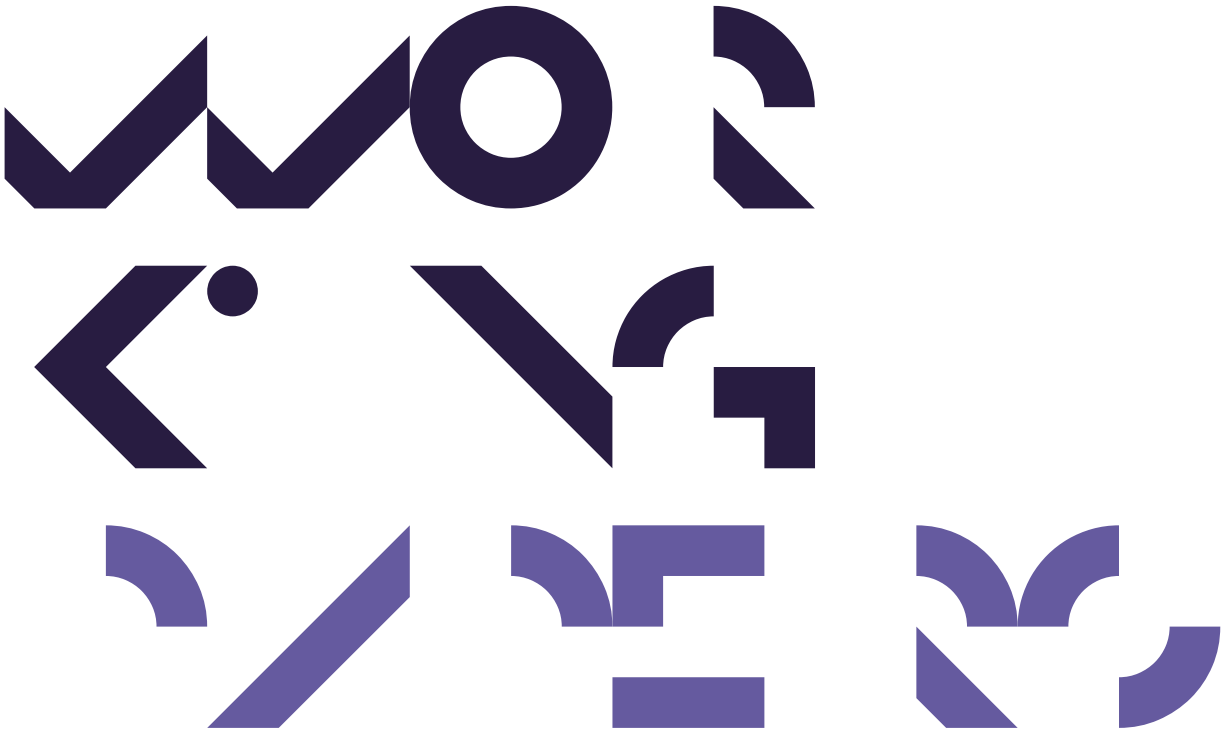


Economics Working Papers

2014-25

The Danish National Tests – A Practical Guide

Louise Voldby Beuchert and Anne Brink Nandrup

AARHUS
UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

The Danish National Tests – A Practical Guide

Louise Voldby Beuchert
Aarhus University
Dept. of Economics and Business
lbeuchert@econ.au.dk

Anne Brink Nandrup
Aarhus University
Dept. of Economics and Business
anandrup@econ.au.dk

August 2015, preliminary

Abstract

In 2010, the Danish National Tests were implemented in the public compulsory schools as a mean of evaluating the performance of the public school system. The extensive test program consists of ten mandatory tests in six subjects in grades 2 through 8. In this paper, we share our insights from working with the first four rounds of the test data. We provide a brief introduction to adaptive testing, the available data and general data issues including missing data, test participation and data transformations. Additionally, we construct a standardized measure of the raw test results within each test and argue that this is often a more feasible measure for data analyses compared to the transformed test score presented to pupils and teachers. We provide the reader with preliminary analyses of the relation between pupils' national test results and a wide range of pupil background characteristics as well as pupils' 9th grade examination marks. We document a stable test score gap across grade levels and socio economic background and discuss the prospects of the national test data for future research.

JEL classification: I20, I21, I24

Keywords: Test scores; Adaptive testing; Test score gap

Acknowledgements: Financial support from TrygFonden's Centre for Child Research is gratefully acknowledged. We are thankful for constructive comments and suggestions from researchers and seminar participants at the TrygFonden's Centre for Child Research. The usual disclaimers apply.

TRYGFONDENS
BØRNEFORSKNINGSCENTER

1. Introduction

In 2010, the Danish government introduced a yearly national testing scheme, covering six subjects to the public schools. This test program is called The National Tests. The national tests are mandatory and consecutively test pupils from grade 2 through grade 8. Thus, they offer a unique opportunity of following the academic achievement of public school pupils throughout compulsory school. In this paper, we collect and share our insights from working with the first four rounds of the national tests. Further, we provide the reader with preliminary analyses of the relation between pupils' national test results, student background characteristics and 9th grade exit exam results. This is meant as a starting point for discussing and developing new research ideas exploiting the national test data.

The national tests were introduced to the public primary and lower secondary school in the school year of 2006/2007 with the purpose of contributing to the continuous evaluation and improvement of the public school system. Establishing a reliable evaluation culture in the public schools was one of the main recommendations in the 2004 OECD review of the public education policies in Denmark (OECD 2004). However, the 2007 test evaluation indicated severe problems with the test content. After redevelopment, pilot testing and trial runs, the national tests were officially launched in the beginning of 2010. The tests are still in their implementation phase, as only a few years of test results are available, and the tests are still monitored for possible data problems (Undervisningsministeriet 2012a, Rambøll 2013). But while the test program is now an integral part of public schooling, it is still heavily debated among both researchers and practitioners.

When considering the national test data, two main questions are raised: How is the student skill level estimated and what does it measure? To answer these questions, this paper briefly introduces adaptive testing in general and the national tests in particular. This run-through allows us to summarize potential pitfalls and advantages of the national tests. Here, we have collected information regarding the nature of the program from, among other, The Danish Ministry of Education as well as two evaluation reports published in 2007 and 2013, respectively. Secondly, this paper attempts to shed light on the relations between the tested skills and later, more common measures of success. For this, we use pupils' 9th grade exit exam results. Overall, we find that the national tests are able to measure skills that are at least very highly correlated with the skills measured by the exit examination marks. However, the correlation between examination marks and test results are significantly less for pupils of non-Western backgrounds.

Additionally, we investigate general data issues and provide descriptive analyses of the test results obtained in the period 2010-2013. We base our analyses on a standardized measure of the raw test results within each test and argue that this may often be more feasible for analysis compared to the transformed test score presented to teachers, parents and pupils. Not surprisingly, we find that low socioeconomic status is generally related to lower test results. Also, girls' reading scores are on average significantly better compared

to boys', however, this patterns is reversed for math based subjects, in particular physics and geography. Being a non-Western immigrant or a descendent hereof is associated with significantly lower test results in all tests, yet this effect is significantly less negative for English scores. Interestingly though, when investigating how well test scores in previous grades and student background characteristics explain student achievement in later grades, student background characteristics do not increase explanation power once controlling for the test results from earlier grades. Further, we document a stable test score gap across grade levels among pupils from different backgrounds.

We emphasize that the empirical findings of this paper are purely descriptive. Still, we find it relevant to include these as a practical introduction to the national test data.

The structure of the paper is as follows; Section 2 gives a broad introduction to the national tests and the adaptive testing process used to determine the skill level estimates. In section 3 we describe the underlying theoretical model for adaptive testing, the Rasch model. Available data and general data issues are documented and discussed in section 4, followed by empirical analyses in section 5. Finally, section 6 concludes with recommendations for data analysis and prospects for future research exploiting the national test results.

2. The Danish National Tests

This section describes the background and content of the national test program – specifically how this estimate of pupil skill[1] is found by adaptive testing and the implications of this. The test results as measured by the national tests are not comparable to the results of regular linear test results. Rather, they are supposed to be a very precise estimate of the skill level within the specific cognitive area tested.

2.1 Background and content

By *Folkeskoleloven* (The Public School Act) §13, stk. 3, all children enrolled in Danish public schools must take ten national tests during compulsory schooling.[2] The respective grade and subject of each test is presented in Table 2.1. The table illustrates how pupils are subject to a reading test every second year, a math test in grades 3 and 6, and finally other subject-specific tests in grades 7 or 8. The reading tests in Danish as second language are voluntary on the school basis[3]. The main purpose of the national tests is for the teacher to gain insight into the individual achievements of the pupils, and use this insight as an evaluation tool when forming the individually targeted teaching plans[4] (Skolestyrelsen 2010a). The individual test results are

[1] The test results are termed *estimated pupil ability* (in Danish: estimeret elevdygtighed), but to avoid confusion with the general literature, where ability typically denotes time-invariant inherited skills, we name test results by estimated pupil skill level.
[2] Private schools are currently exempted from test taking; however, it is possible for them to participate on a voluntary basis. These will not be discussed here.
[3] These tests were first introduced in the fall of 2012 as a result of a pilot project in June 2012.
[4] Teachers are required to regularly compose individually targeted student plans (in Danish: Elevplaner) that serve to increase the focus on learning progress and as a communication tool between teachers, pupils and parents (Undervisningsministeriet 2010).

confidential and only known by the pupil, his subject-specific teacher, and his parents. However, parents are only presented with the test result on a crude five-point scale. Through the online test system teachers are able to access test scores on a more detailed scale from 0 to 100. Here, they are also able to recall the answered questions as well as the sequence of answers for a given pupil in their subject specific class[5]. This information can then be used to target teaching for the individual student. Secondarily, the results of the national tests are meant as a monitoring and policy device for the school principal, school board, and municipality authorities. Because the individual test scores are confidential, only mean test scores on the cohort-level is available to the school board and municipality authorities, while the overall national distribution is presented to the public (Skolestyrelsen 2010b).

The tests are mandatory and completed in the period January through April, i.e. at the end of the school year. The test period is pre-defined and may change from year to year.[6] Additionally, schools may choose to repeat the same grade level test. These voluntary tests are conducted during the autumn one year before, in the same year, and/or one year after the respective grade of the test (represented by the shaded areas of Table 2.1). The voluntary tests draw from the same item bank as the mandatory ones.

Three separate cognitive areas within each subject are tested simultaneously. These cognitive areas are called ***profile areas*** and are listed in Table 2.2. For example, the tests in Danish, reading evaluate pupil skills within the following three profile areas: language comprehension, decoding, and reading comprehension. The skill level is estimated separately for each profile area of the subject. This will be carefully described in the following sections.

Table 2.1 Overview of 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					X		X		

Notes. X indicates the grade levels subject to the national tests. The test for Danish as second language is marked with red as these are only voluntary. Additional to the mandatory tests, it is possible to test pupils in the grade level above or below on a voluntary basis. These are illustrated by the shaded areas.

[5] The test items are confidential and may not be leaked to the public.

[6] The mandatory test period of 2010 was planned in February 15 – April 30, with a retesting period for sick absentees of June 7 - 25. In 2011, the mandatory test period was planned in January 10 – April 29 and a sick period of June 6-24. For 2012 the dates were January 16 – April 30 and May 29 – June 11 and for 2013 January 21 – April 30 as well as May 27 – June 28 (this year's retesting period was augmented to incorporate pupils affected by a nationwide lockout of teachers).

2.2 Test properties

The national tests are IT-based, objective, and adaptive. IT-based simply means that the test is performed online with each pupil sitting by a computer. To test the pupils, the subject-specific teacher has to pre-book a test session within the test period. During a test session, all pupils of a given class are placed in the same classroom with individual computers. Each pupil then uses his or her unique login to log on to the online national test home page. The test results are subsequently saved in a personal electronic profile.

Objectivity arises as no teacher assessment is required, i.e. the test system chooses the questions and calculates the test results. Previous studies have shown indications of teacher bias in the evaluation of pupil abilities, i.e. that characteristics, such as gender and race, significantly affect teacher perceptions of pupil performance (e.g. Dee 2007, Downey and Pribesh 2004). Of course the objectivity of the test results relies heavily on the test questions being objective across gender, race, residential area etc. This issue is considered in detail in Section 2.3.

Table 2.2 Subject specific profile areas of the national tests

Subject	Profile Area 1 (p1)	Profile Area 2 (p2)	Profile Area 3 (p3)
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)	Living organisms' interactions (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. *Source:* The Danish Ministry of Education (Skolestyrelsen 2010b).

Simplified, adaptive testing means that the pupil is presented with questions (called items) of different difficulty levels based on whether or not he was able to answer the previous question. Here, it does not matter how many questions the pupil is able to answer correct as opposed to regular linear tests. Instead the difficulty level of the question answered is of importance. The test result is an estimate of the pupil skill level. As the difficulty level of the items is continuously updated, this roughly corresponds to the item difficulty of the final question and is argued to be a more precise and detailed estimate of a pupil's skill level compared to what can be revealed by linear tests (Review 2007). To ease interpretation of the test results, teachers and parents are presented with a transformed test result on a scale from 1 to 100 and 1 to 5, respectively.

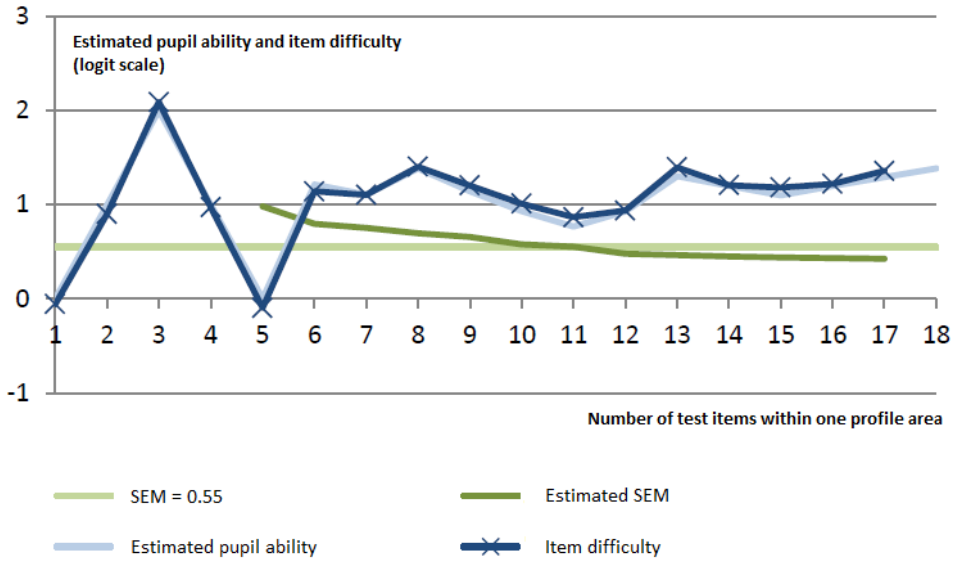
As briefly explained above, the adaptiveness of the national tests implies that the difficulty level is objectively updated and the corresponding skill level estimated as a function of the individual pupil's level of proficiency during the test, until a given stopping rule is met. This is graphically illustrated in Figure 2.1 with each *x* marking the items asked within a single profile area. The light-blue line illustrates the estimated skill level and the dark-blue line illustrates the difficulty level of the items. Thus, the test result is then based on only two parameters; the difficulty level of the question and the true skill level of the pupil.

We start by considering the item difficulty: The difficulty level of a test item denotes the relative level of difficulty when all test items within a profile area are ranked on the continuous logit scale on the [-7; 7] interval (UNI-C 2012a). In practice, the difficulty level is determined by a test pilot of approximately 700 pupils; the answers are subsequently evaluated by a Rasch analysis in order to determine if the item measures the intended area and on which difficulty level (Rambøll 2013). In the test situation, the first item presented to a pupil within each profile area is designed to have difficulty level 0 (corresponding to around average). As illustrated in Figure 2.1 the next four items within the same profile area are chosen based on the pupil's answer to the previous item – but without explicitly estimating the skill level. Thus, a correct (wrong) answer triggers the next item of that profile area to be of a difficulty level of approximately 1 above (below) the previous level (UNI-C 2012a).[7] Hence, the difficulty level of the test items is very volatile in the beginning of a test period. Critics of the national tests have voiced their concerns that this causes some, particularly skittish pupils to be 'trapped' at too low initial difficulty levels, because a wrong answer is punished relatively harder in the beginning. From the sixth item and onwards the item difficulty level is based on the (updated) estimated pupil skills (see Figure 2.1, the item difficulty and estimated skill level no longer coincides). Thus, the pupil is given an item that is approximately of the same difficulty level as the estimated pupil skills based on the sequence of items already answered. This Rasch algorithm implies that a pupil

[7] From the school year 2014/15 the run-in period is adjusted to include three instead of five items and the difficulty level of item two and three will only change by +/- 0.5 logits depending on a correct answer. Further, the difficulty level of the first item now accommodates the mean difficulty level within specific profile areas and, thus, not 0 in all cases (Undervisningsministeriet, 2015 January). Please consult the Ministry of Education for the latest description of the adaptive algorithm.

should be given questions with equal probability of a correct/false answer. For a more detailed description of the Rasch model, adaptive testing, underlying assumptions and estimation, see section 3, Rasch (1960) or Andersen (2002).

Figure 2.1 A typical test example for a single profile area from UNI-C 2012a (Figure 4)



Notes. The figure illustrates how the estimated pupil skill level, item difficulty and estimated standard error of measurement (SEM) may progress during the test of a single profile area. *Source*: UNI-C (2012a).

The process continues at least until the skill estimate satisfies a standard error of measurement (SEM) below 0.55.[8] [9] In short, the SEM denotes the variation in the pupil's ability to correctly answer the items. As the difficulty level of the test item is distributed on the [-7; 7] range logit scale only very few pupils will answer all questions consistently correct (or incorrect).

The SEM is illustrated with the green lines in Figure 2.1. Like the skill level, the SEM is also (re)estimated after each item answered following the 5th item; see the dark-green line. In the example the pupil starts out with a SEM of 1 and reaches an estimated SEM of 0.55 around the 11th test item. By answering more items the SEM is further reduced and the estimated skill level (light-blue line) is converging to a skill level around 1.3. In practice, test results reported to the teacher during the test session are marked with a color (green, yellow, red)[10] depending on the level of the SEM. The teacher can then monitor when the pupil has answered

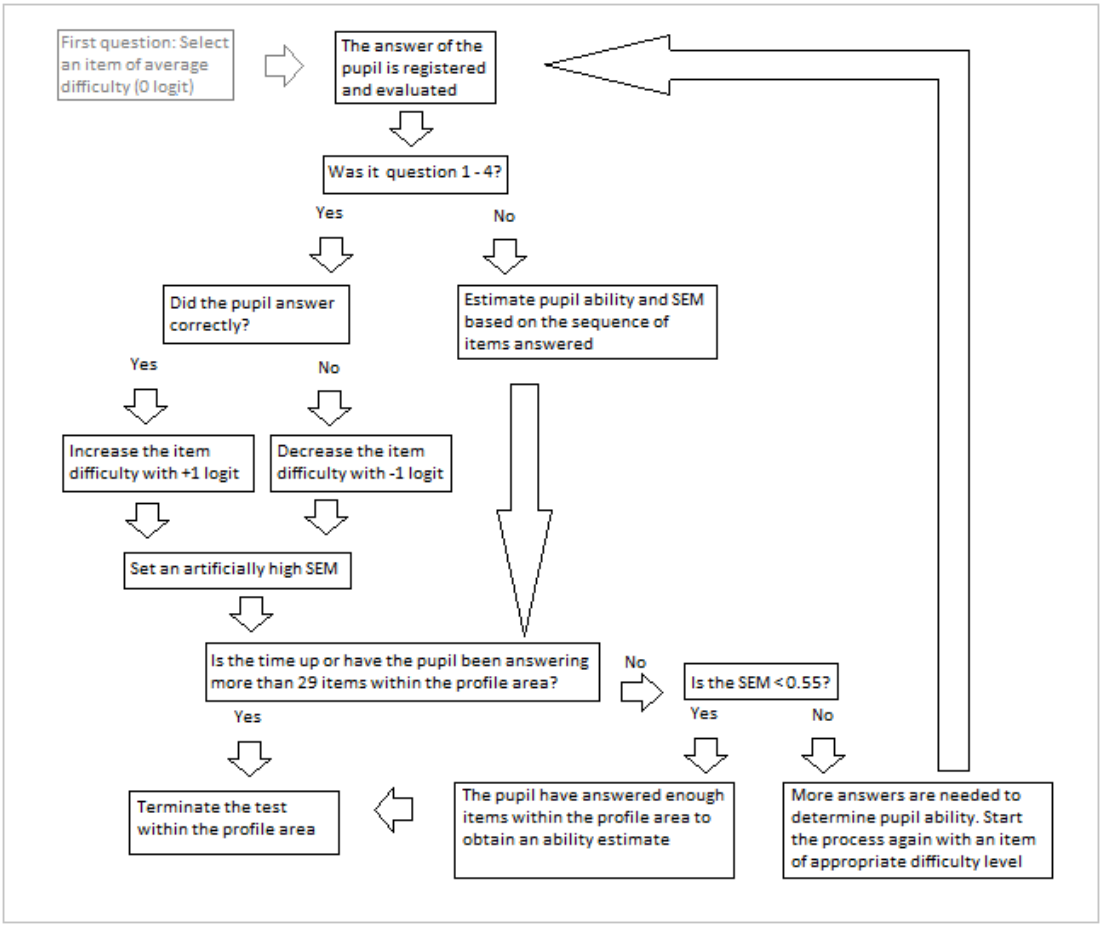
[8] SEM $= 1/\sqrt{s^2}$, where s^2 is the sum of the variances of the items that the pupil has attempted to answer (see UNI-C 2012a, appendix 1). Originally, a SEM below 0.3 was the limit of sufficiently precisely estimated parameters. In practice, the statistical uncertainty of test results are substantially larger (0.55) compared to the intended 0.3 (Undervisningsministeriet 2014).

[9] The test result has a 5% confidence interval of $\pm 2 \cdot SEM$. Unfortunately, UNI-C is currently not able to retrieve and provide information about the SEM. With time, it is expected to be added to the available data. The same applies to information about the duration of the specific tests.

[10] The estimates are marked red for less than 5 items answered within a profile area, yellow for 5-30 items within a profile area and a SEM above 0.55, and green for a SEM below 0.55 or more than 30 questions answered within a profile area (the latter is primarily the case for very high- or low-performing pupils).

sufficiently many items and may choose to terminate the test session[11]. The details of the testing process within a single profile area are summarized in Figure 2.2.

Figure 2.2 The testing process of a single profile area



Notes. The figure summarizes the test cycle of a single profile area. The difficulty levels of the first five questions are special in the sense that they are not based on estimated skill level. The first task is by construction of approximately average difficulty level. Hereafter, the difficulty levels are adjusted according the answers of the pupil. Questions 2-5 are increased (decreased) by 1 logit depending on a correct (wrong) answer and from question 6 and onwards the estimated skill level is the basis of the selected difficulty. *Source*: UNI-C (2012a).

Incorporating the three separate profile areas within each subject into the testing process, the pupil has 45 minutes to answer as many items as possible within the subject. The pupil is alternately presented with questions from each of the three profile areas until the skill level estimate within one profile area has reached a SEM below 0.55 (usually within 15 items). Hereafter, the items alternate only between the remaining two profile areas, etc. When the SEM of all three profile areas are below 0.55, the items once again alternate between all three profile areas for the remaining period of time, to further reduce uncertainty of the skill level estimates.

[11] It is recommended by the Ministry of Education to continue testing throughout the booked test session to ensure the lowest possible uncertainty of the test results.

The main drawback of the adaptive tests is that this approach raises the requirements to the test items compared to linear tests, which in turn increases the resources demanded for developing and maintaining the test system (see discussions in Review 2007, and Rambøll 2013). The next section is devoted to this central issue on the properties of the test items.

2.3 Test items

Any test item must satisfy the Rasch requirements in order to contribute to correctly estimating the pupil skill level (Andersen 2002). Namely, two pupils of *equal* intelligence must have the *same* probability of correctly answering a given item, independently of gender, socioeconomic background etc. (also known as the criterion of homogeneity). As such, the comparison of two pupils should be independent of items drawn from the test item bank. Of course the polytomous questions[12] then require that all sub-questions of a given task are of exactly the same difficulty level. If the Rasch requirements of invariance of comparisons are not met, then using the total score to characterize the pupil's skill level is not justified. A Rasch analysis tests the fit between the data and the underlying Rasch model.

In practice, the items are formed as multiple choice questions but with a varying number of options. According to UNI-C (2012) two thirds of the items are dichotomous, i.e. one can either answer correctly or not, while one third is polytomous. This is called the *complexity* of the test question. The first five items within each profile during a test are all dichotomous (a complexity of one). Only thereafter it is possible to get a polytomous question. Available test items fall into as much as 15 categories (Rambøll 2013): e.g. multiple-choice, (word) insertion, word splitting, and coloring-tasks. Multiple-choice items comprise the majority of the test items. Small steps have been taken to incorporate a time dimension into the reading items concerning decoding (profile area 2). So pupils, who are already able to read, can have their reading speed tested. This feature is, however, not expected in the near future.

The test result as registered by the system may change over time (Pøhler and Sørensen 2010). E.g. if the teacher in 4th grade decides to have a look at a pupil's 2nd grade test scores, they may have changed, as a consequence of the continuously updated item bank. Items that are subsequently removed from the bank will then not affect the 2nd grade test result. This supposedly ensures that test results are always comparable across cohorts. In reality, the test results would only differ slightly, but some pupils on the margin or pupils, who only just satisfy the stopping rules, may be affected more by an update. School or class averages on a specific grade level in two different years may be more comparable when test items are the same across the observations. However, by that reasoning new items should also be excluded for comparability of test results. Also, all test items must pass the Rasch analysis to be included in the national tests in the first place, so one has to assume that even replaced test items indeed *did* satisfy the Rasch requirements. Consequently,

[12] Questions containing several subquestions each of which has right or wrong answers. I.e. one can answer e.g. 50% correct.

it is unclear why removing an item from the bank without updating the test results would cause incomparability of skill level estimates.

There is no public information of the yearly item turnover rate in the item banks, when replacements are executed or even of the current size of the item bank.

In the next section, we summarize the main points of criticism concerning adaptive testing and the Danish national test scheme raised by the 2007 and 2013 test evaluations as well as other debaters.

2.4 Caveats and practicalities

The national tests have been designed as an adaptive test system among other things to be able to capture the wide range of pupil skills present in a typical Danish class (Wandall 2011). Regular linear tests would require a substantial amount of test questions to be able to precisely determine pupil skill level – particularly, concerning the top and bottom ranked pupils. From an analytical point of view, test items that are either too easy or too difficult for a given pupil reveal very little about the child's true abilities. However, with questions that target the level of the individual pupil based on answers already given, IT-based adaptive tests should objectively and precisely estimate the skill level of a child within the specific cognitive area of testing. For example, the test items for profile area one are designed only to evaluate skills of this area and not to measure skills of the remaining two profile areas within the respective subject (Review 2007). Of course, some skills are not possible to evaluate with IT-based tests, for example, writing skills or independent oral performance. But for the profile areas listed in Table 2.2, a precise skill level estimate should be found. However, other caveats both theoretically and practically have been raised concerning the national test program. Below we will discuss the four main caveats in turn.

First of all, the validity of the testing program hinges on whether the test items satisfy the properties of the Rasch model. In the 2007 evaluation, an independent review panel was asked to evaluate the 2007 pilot of the national tests, see Review (2007). The review panel concluded that the presented questions satisfied the properties of the Rasch-model to high standards (compared to international standards). However, the number of available questions was too few, and the polytomous items revealed some local item-dependence. Therefore a special focus on the polytomous items was recommend for the following process of developing and monitoring the item bank. A recently published test evaluation (Rambøll 2013) still outlines certain difficulties concerning the number of items in the test bank. Particularly, there is a shortage of items in the top 10 percent difficulty levels. This implies that high-skilled pupils may exhaust the items of the relevant difficulty levels, in particular when voluntary tests are used in combination with the mandatory tests. The two evaluations confirm that it is a lengthy process to develop an appropriate and adequate item bank. However, the evaluators also note, that despite the challenges with developing a sufficiently large item bank, the adaptive principle implies that it takes approximately 50 percent *less* test questions to determine pupil

proficiency compared to regular linear tests. Therefore adaptive testing is still recommended despite the resource demanding process maintaining the item bank[13].

Secondly, due to the adaptiveness of the test, for some children it only takes 10 questions to precisely estimate the skill level, while for others it may take more questions. A concern is that this may influence the practical test session and cause some pupils to abort the test before a sufficient SEM is reached. By default a test within one subject (three profile areas) is terminated after 45 minutes, but the teacher may prolong the test until 180 minutes – or stop it at any time (note that the SEM has to be less than 0.55 or the pupil has to have answered more than 29 items within all profile areas to validate the test result). Based on the structure of the tests, there is no reason to suspect that pupils are forced to randomly select answers in the end of the test in order to finish all questions 'in time'. However, it may be of concern that the teacher can prematurely terminate the tests, so pupils may be preoccupied with just finishing in order to get on with other things. Indeed Rambøll (2013) notes that some children, especially in the younger grades, may not be entirely focused on answering the test items rather than finishing quickly in order to go out and play.

On the other hand, the teacher has the option of discontinuing the tests for specific pupils if he believes that the pupil is no longer performing optimally. In effect this means, that pupils with concentration difficulties may only take the test in short intervals with breaks of up to several days in between. The reasoning behind this reads: a test that does not take the pupils' difficulties into consideration is misleading (Pøhler and Sørensen 2010). By the same reasoning teachers are allowed to provide pupils with test aids based on their everyday assessment of the pupils' needs. Because a pupil is furthermore only presented with tasks that match their proficiency level the adaptive national tests also allow pupils with special education needs to be tested, (Skolestyrelsen, 2010a). Thus, learning disabilities or physical disabilities should not affect the participation rate of pupils. E.g. there is a series of test items that can be answered without the use of a computer mouse. The test conditions are noted in the student plan in order to evaluate the optimal performance of the pupil. Unfortunately, it is currently not possible to obtain information of the duration of the individual tests or the use of disconnections and test aids. Note that from the school year 2014/15, there are some changes to the national test program. These changes include criterion-referenced test results (in addition to the above described norm-referenced test results) and new guidelines for timing of the test-session and provision of test aids as well as modifications of the run-in period (see footnote 7).

Thirdly, because the test items are of a multiple choice-form there is a possibility of guessing the correct answer without understanding the question. I.e. the skills of weak pupils are potentially overestimated. Based on the following four arguments, Kreiner and Wandall (2012)[14] argue that this is not a concern: i) To be able to guess perfectly random is a very complex strategy that especially the weaker pupils are unable to master,

[13] See Appendix 4 of Rambøll (2013) for a thorough review of pros and cons in relation to the adaptive principle.
[14] Both authors were part of the developing committee for the national test program.

ii) The methods used to score the tests incorporate a certain degree of randomness, as in case of qualified guessing, iii) The initial testing of the items has revealed and subsequently excluded items in which weak pupils have a surprisingly high rate of success and iv) Based on previous results, the weaker pupils in general seem to have a significantly lower rate of success compared to what they would have obtained, should they have chosen to guess systematically. The latter issue arises both when weaker pupils are unable to guess randomly but also if these leave blank items instead of risking a wrong answer when they are uncertain of the correct one. When scoring the tests a blank answer is considered a wrong answer.

The adaptive function ensures in principle[15] that the strongest pupils are continuously challenged as they are confronted with increasingly difficult tasks until they fail to answer correctly. Many of the high-skilled pupils are not used to being challenged in ordinary proficiency tests, which may distort their motivation during test taking. At the same rate, weak pupils do not experience the failure of not being able to answer more than a few questions.

Finally, there is the concern of how the test results are actually being assessed and used in the class room teaching. Specifically, if the national tests are used as an evaluation tool for the individual learning, as they were intended to, or if teachers are more concerned with *teaching to the test*. With respect to the application of the tests, it seems that the majority of both teachers and principals employ the tests pedagogically and managerially (Rambøll 2013). But, the information is primarily used summatively whereas the intended formative application seems rare. Interestingly, the Rambøll (2013) evaluation of the national tests points in the direction of a positive effect of the tests on the overall pupil proficiency. And more importantly, this does not seem to be driven solely by teaching to the test-effects or the like. Furthermore, the introduction of the test scheme may have contributed to strengthening the evaluation culture at the schools. This is, however, not regarded as a result of the tests alone rather than the combination of the tests *and* other evaluation tools guided by a strong management team (Rambøll 2013). Also, the evaluation committee notes that the pupils generally view the tests positively.

2.5 Reporting the test results

There are three means of test score reporting, whereof data provided by UNI-C contains two measures: The estimated skill level on a logit scale (*theta*) within each profile area and a recalculated test score on a 1-100 scale (*point*) (see subsection 3.1.2 for details of the transformation, note that the 1-100 scale *does not* correspond to actual percentiles).

Based on the latter, UNI-C reports the test results of pupils to teachers and parents on a five-point scale explained below. In particular, parents to tested pupils receive a short letter in which achievement of their

[15] Not considering the shortage of very difficult questions (Rambøll 2013).

child is explained within each profile area in text form (see Wandall 2011, Appendix 2 for an example). The five-group scale reads[16]:

1. Considerably below average (points: 1-10)
2. Below average (points: 11-35)
3. Average (points: 36-65)
4. Above average (points: 66-90)
5. Considerably above average (points: 91-100)

Additionally, the teacher has access to test scores in points for each pupil in the relevant profile areas as well as overall ratings per pupil (raw average across profile areas) and class averages.

3. The Underlying Model and the Test Reportings

3.1 Adaptive tests and the Rasch model

An adaptive test can be based on a Rasch model (Andersen 2002) as is the case of the Danish national tests[17]. In the standard Rasch model, the probability of a correct answer is a function of pupil skills and the difficulty of the test alone, where the difficulty level of the items and the skill level are measured at the same logit scale. If the items satisfy the properties of the Rasch model, then the estimate of pupil skills is valid, reliable and objective (Rasch 1960).[18] The items are optimal if the difficulty level matches skill level, i.e. the probability of a correct answer is equal to one half given only the skills of the pupil.

3.1.1 The Rasch model and the skill level estimates

The Rasch analysis is valid given three assumptions. Firstly, skills are assumed to be a unidimensional trait. Secondly, the earlier mentioned requirements of invariance of comparisons (local independence of items), and thirdly, that the response function of a pupil can be characterized by a Rasch response function.

For dichotomous items the Rasch model is given by a logistic function:

$$\Pr(X_{ni} = 1|\theta_n) = \frac{\exp(\theta_n + b_i)}{1 + \exp(\theta_n + b_i)}$$

Where $X_{ni} = 1$ denotes a correct response, θ_n denotes the true skill level of pupil n and b_i is the difficulty of item i.

Recall that only around two thirds of the test items are dichotomous. The last third contains several subquestions that all contribute to the assessment of whether the item is answered correctly, partly correct or

[16] This distribution roughly corresponds to the expected distribution of passed examination marks, 2, 4, 7, 10 and 12, of the 7-point grade scale (Pøhler and Sørensen 2010, and Wandall 2011). It is also possible to obtain the failing marks -03 and 00.

[17] In general item response theory (IRT) denotes the paradigm for the design, analysis and scoring of tests, where the varying difficulty of each item is incorporated into the scaling. Some psychometricians consider the Rasch model to be a one-parameter IRT model, while others consider it a completely different approach.

[18] The Journal of Applied Testing Technology have recently (2011) published a collection of articles on the Rasch model and how to analyze if an adaptive test is satisfied including testing the item properties.

wrong. This raises the requirements to the items, as each question within an item should have exactly the same difficulty level independent of the previous question. The polytomous Rasch model is generally given by (Andersen, 2002):

$$\Pr(X_{nji} = j|\theta_n) = \frac{\exp(\theta_{nj} + \varepsilon_{ij})}{1 + \sum_{q=1}^{m-1} \exp(\theta_{nq} + \varepsilon_{iq})} \quad for \quad j = 1, \dots, m-1$$

And $$\Pr(X_{nji} = m|\theta_n) = \frac{1}{1+\sum_{q=1}^{m-1} \exp(\theta_{nq}+\varepsilon_{iq})}$$

With $i = 1, \dots, k$ items and $j = 1, \dots, m$ response categories of the item of individual n. $\boldsymbol{\varepsilon}$ denotes the difficulty level. The number of response categories corresponds to the complexity of a test item. Also, note that the dichotomous Rasch model is a special case of the polytomous with $m = 1$.

The skill level is iteratively determined by the Newton-Raphson method. The test result is measured with 10 decimal places and can be considered continuous.

3.1.2 Skill level estimates and the point scale

For easier interpretation the test results are transformed by authorities to be presented on a point scale from 1-100. For this a sigmoid function (an S-shaped function) is used. To transform the distribution of any variable into percentiles one would usually use the empirical cumulative distribution function of the distribution in question as the sigmoid function. The point transformations of the national tests are, however, obtained using a slightly different cumulative distribution and hence cannot validly be interpreted as percentiles[19].

Based on a specified number of obtained test results within each profile area in 2010[20] a sigmoid function has been fitted to the empirical CDF. The slope of the sigmoid function is allowed to differ above and below the median of the sample:

$$g(\hat{\theta}_n) = \begin{cases} \dfrac{100}{1 + exp(-e(\hat{\theta}_n - f))} & for \ \hat{\theta}_n < f \\ \dfrac{100}{1 + exp(-k(\hat{\theta}_n - f))} & for \ \hat{\theta}_n \geq f \end{cases} \qquad (1)$$

where $\hat{\theta}_n$ denotes the estimated skill level for individual n, $g(\hat{\theta}_n)$ is the transformed point score, f is the median and e and k is the slope of the curve just below and above the median, respectively. See Appendix A for parameter values across profile areas. Point scores are always rounded up to the nearest integer.[21]

[19] In the official literature, the point scale is often referred to as "the percentile scale" (e.g. UNI-C 2012a). We will refrain from using the term percentile scale in this paper.

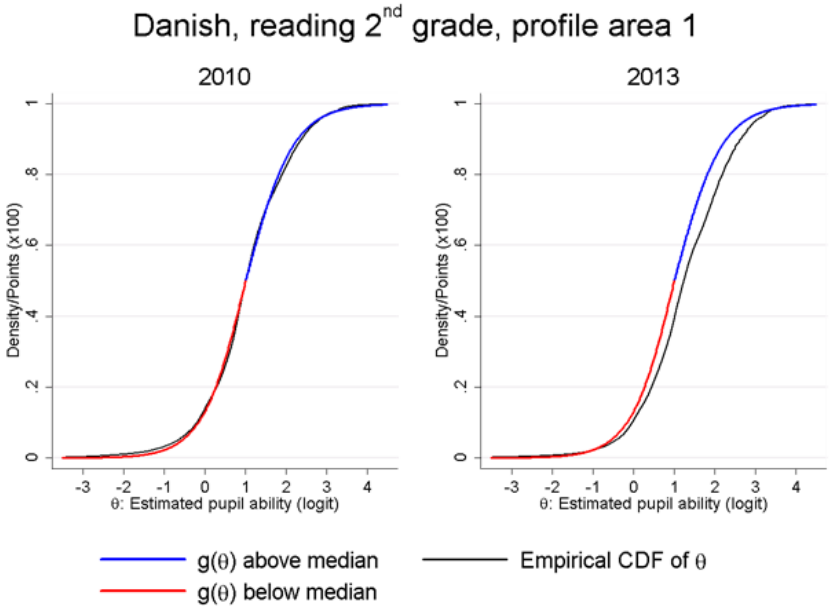
[20] Wandall (2011) states that the percentile transformation is based on the test results from the first three weeks of full-scale testing in 2010, however these numbers do not add up to the number of estimates used according to UNI-C (see Appendix A).

[21] We are able to replicate all but 515 (<0.03%) transformations from θ to point of the 2010-2013 tests. Of these 515 observations, the majority of the differences amount to ±1 point but overall they vary between -28 and +12 points.

Figure 3.1 demonstrates how $g(\hat{\theta}_n)$ compares to the empirical CDF of the test results within profile area 1 of Danish, reading 2nd grade. As expected, $g(\hat{\theta}_n)$ resembles the empirical CDF for a wide range of test results in 2010 (left-hand panel). The slight differences illustrate where the point scores differ from the empirical percentiles. In 2013, however, the distribution of test results have been shifted to the right and no longer coincides with $g(\hat{\theta}_n)$. Consequently, the point score is considerably larger than the empirical percentiles for pupils in the middle of the skill level distribution. For example, an estimated skill level of 1 logit still corresponds to 50 points even though the skill level estimate lies around the 40th percentile (the tails are still largely coinciding). As such, the point score always provide a test result relative to pupil skill level in 2010. But even for 2010 one cannot interpret points as empirical percentiles. Moreover, they do not say much about where the pupil belongs in the skill distribution compared to his/her peers.

The main lesson to learn from this section is that while the logit scale of the test results may seem unfamiliar; this measure is considerably more detailed and will often lead to easier interpretations of coefficients when used in analysis. However, the distributions of the skill level as measured by the tests vary considerably across tests, years and profile areas (Appendix B). Hence, in the remainder of this paper we use a standardized measure of the raw skill level estimates; see Appendix C for details about the standardization procedure.

Figure 3.1 Transformation of the test result to the linear point scale



Notes. The figure illustrates the relation between the test result and the percentile score presented to the teachers for profile area 1 in the test Danish, reading 2nd grade. The actual cumulative distribution of test results is also shown. The shape of the curve differs across tests and profile areas.

4. Data

This section presents available data of the 2010-2013 mandatory national tests. Data is provided by UNI-C (The Danish Agency for IT and Learning).

4.1 Documentation

Table 4.1 describes the key variables in the national test data. For each test observation, the subject of the test (FAG), the test identification number (FAGID) and the date and time of the test (TESTTID) is available. The skill level as estimated by the tests (THETA) and the point score (POINT) are available for each profile areas separately and are denoted by _P1, _P2 and _P3, respectively. See Appendix B for descriptive statistics of the raw skill level estimate (THETA) by year, test and profile area. Each test observation contains the pupil's anonymized civil registration number (PNR) enabling us to link pupils' test results to all other register data maintained by Statistics Denmark. School identifier (INSTNR), grade level (KLASSETRIN), school type (SKOLETYPE) and municipality identifier (KOMMUNENR) may also be available, but these are not obtained directly from the test system. Rather, they are matched from school registers at the time of data extraction and, for this reason, they can be highly unreliable. For example, a pupil who is tested in April 2010 at school *x*, but is transferred to school *y* before the time of data extraction, will appear to attend school *y* in the test data. Likewise, school types that are not subject to the mandatory tests may appear.

Table 4.1 Documentation of variables

Variable name	Variable description
PNR	Encrypted civil registration number (personnummer)
TESTTID	Timeslot for test taking (Tidspunkt for afholdelse af test). Note, that this is denoted per hour (likely the beginning of the booking time slot) for all data sets except the mandatory tests of 2012, where the individual end time of the test is likely registered.
FAG	Subject and grade level of the test (Fag og klassetrin)
INSTNR	Institution number of the pupil at the date of data extraction (*not* at the time of the test!)
KOMMUNENR	Municipality number (of the school)
SKOLETYPE	Corresponds to the variable INST2 maintained by Statistics Denmark¤: 121= Elementary school (grundskole) 126= Special schools for children (Specialskoler for børn) 129= Treatment facility options/homes (Dagbehandlingstilbud/hjem)
KLASSETRIN	Grade level (Klassetrinnet)
POINT_P1	Point score for profile area 1 on a percentile scale (Opnået resultat i profilområde 1 på percentil skala) [1;100]
POINT_P2	Point score for profile area 2 on a percentile scale (Opnået resultat i profilområde 2 på percentil skala) [1;100]
POINT_P3	Point score for profile area 3 on a percentile scale (Opnået resultat i profilområde 3 på percentil skala) [1;100]
THETA_P1	Estimated skill level for profile area 1 on a logit scale (Elevdygtigheden i profilområde 1 målt

	på logitskala) $(-\infty;\infty)$[22]
THETA_P2	Estimated skill level for profile area 2 on a logit scale (Elevdygtigheden i profilområde 2 målt på logitskala) $(-\infty;\infty)$
THETA_P3	Estimated skill level for profile area 3 on a logit scale (Elevdygtigheden i profilområde 3 målt på logitskala) $(-\infty;\infty)$
KOMMUNE_NAVN	Name of the municipality (of the school)
FAGID	Test identification number (see Table 4.2) – only included in the test data of 2010 and 2011

Notes. The table summarizes the variables present in the test data. The variable FAGID is added by the authors. *Source*: UNI-C, data documentation 2011. ¤ From 2012 the variable SKOLETYPE corresponds to the variable INST3 maintained by Statistics Denmark.

The estimated skill level and point scores are easily identifiable by the FAGID-variable characterized in Table 4.2. The variable FAGID is included in the test data for 2010 and 2011, but must be manually constructed for 2012 and 2013. It is constructed as a three-digit variable identifying the subject specific tests (see also Table 4.2), where the Danish reading tests are identified by 1xx, math tests by 2xx etc. The end digit characterizes the grade level of a specific test, and as such xx4 denotes a test in grade 4.

Table 4.2 An outline of the test identification variable FAGID

FAGID	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Danish/reading	102		104		106		108
Mathematics		203			206		
English						607	
Geography							708
Physics/Chemistry							308
Biology							408
Danish as second language				505		507	

Notes. The table summarizes how each test can be identified by the three-digit variable FAGID.

4.2 Validity of data

This section focusses on the validity of the test data in terms of missing or misreported data. As the national tests are mandatory in all public schools subject to the Public School Act, note that both pupils enrolled in regular public schools and segregated public special education schools are subject to the tests.

Firstly, a few pupils (<0.01%) have multiple test results for the same test. These are to be considered as data errors, e.g. general errors or in consequence of school transfers, misplaced logins etc. More importantly, these are **not** results of pupils failed to satisfy the 0.55 SEM threshold in the first test.

Very few test observations have missing grade levels (KLASSETRIN) and approximately 0.1% of children in the mandatory and sick test data in each year are not tested at the intended grade level. For more than 78% of these observations the grade difference is of plus or minus one grade. Possible explanations for these include pupils being held back a year or skipping a grade, pupils being taught at a different level than their classmates, and misreportings. The pupil identifier is missing for very few test observations in 2012. A

[22] Here, the range of the logit scale is [-7;7]. Distributions of logits generally have thicker tails compared to the normal distribution.

couple of observations (<0.01%) in 2010 and 2012 have invalid encryption of their civil registration number. These invalid identifiers contain letters, and are not possible to match with other data registers.

In the following, only observations of pupils with valid identifiers who are enrolled in mainstream classrooms (normalklasser) in the Danish public schools (i.e. excluding pupils placed in classes for special needs schooling) are included unless otherwise specified.

4.2.1 Missing test scores

The national test program was fully implemented in 2010. However, data reveals a high fraction of missing test results in this year. In particular, a computer crash in March 2010 resulted in the loss of two weeks' worth of testing (see the end of this section). Furthermore, the idea of mandatory testing of all pupils, especially the young pupils, was debated heavily among politicians, teachers, and parents[23]. Correspondingly, the tests faced much skepticism and dissociation from teachers and unions, and may contribute to the fraction of missing test results (Wandall 2011 and Review 2007).

Table 4.3 summarizes the fractions of missing test results each year. Pupils are tested four times in grade 8 compared to e.g. once in grade 7, thus, the missing test observations rather than the missing pupils is described here. Table 4.3 reveals that earlier grade level test and the 7th grade English test seem particularly affected by the 2010 breakdown. A potential cause of the slight increase in missing test results from 2012 to 2013 is the nationwide teacher lockout in April 2013 (elaborated on in the end of this section). In general, the fraction of pupils who are not taking the test is larger for grade 7 and 8 tests and subjects that are not mandatory in the 9th grade exit exam (biology and geography exams are determined by random draws).

A missing test result may have multiple causes. On the individual level pupils may have transferred schools during the school year, have been granted dispensation or simply have chosen to shirk. In Denmark, pupils are registered annually in early September. Should a pupil transfer from a public to a private school, where pupils are not subject to the mandatory national tests before test taking of a given year, the pupil will appear with a missing test result. Shirking, on the other hand, is also a possible cause as the tests are low stake and there are no formal sanctions imposed on pupils (or their schools) who misses them. According to statistics from the Danish Ministry of Education, the fraction of illegitimate absentees in the Danish public school system in 2011/2012 was on average 0.5 percent in the 2nd grade and increasing to 1.4 percent in the 8th grade (UNI-C 2012b). The fractions of sickness absenteeism were 2.7 percent and 3.9 percent, respectively. Compared to these numbers, the missing test fractions seem reasonable, although one should recall that absent pupils can be rebooked in the subsequent retesting periods.

At the school level, a recent or an impending school merger/closure as well as other school specific factors (e.g. school size, managerial perceptions of the test program, etc.) may influence the test rate of the schools.

[23] Besides being a pedagogical principle not to formally grade pupils below the 8th grade level, it is in fact stated by Danish legislation (Folkeskoleloven §13, stk. 5)

There is no evidence of systematically missing test scores in schools that are merged or closed in 2010 or 2011 (Beuchert et al. 2014). Only around 11 schools have participation rates below 80% in 2011 and 2012. In 2012 there is only one school with a participation rate below 50%, this school is merged later in 2012. Also, some 10th grade institutions that additionally offer teaching at the 8th and 9th grade level do not test pupils in some or all years.

Table 4.3 Missing test results by subject and year

Fagid \ Test year	2010	2011	2012	2013
102: Reading, grade 2	14.36%	3.25%	2.18%	3.34%
203: Math, grade 3	12.10%	3.98%	2.31%	3.63%
104: Reading, grade 4	13.75%	3.14%	2.45%	3.12%
106: Reading, grade 6	12.69%	2.93%	2.48%	3.23%
206: Math, grade 6	12.34%	3.23%	2.48%	3.44%
607: English, grade 7	15.94%	5.49%	4.29%	5.83%
108: Reading, grade 8	18.36%	6.14%	4.78%	5.84%
308: Physics/chemistry, grade 8	17.61%	7.36%	6.18%	8.30%
408: Biology, grade 8	18.87%	7.36%	6.30%	7.73%
708: Geography, grade 8	18.20%	7.22%	6.42%	8.13%
Total no.	84,164	26,383	20,847	26,950
(%)	(15.39%)	(4.97%)	(3.96%)	(5.21%)

Notes. The table should be read accordingly: In 2011 3.25% of the 2nd grade cohort was **not** tested in reading. The number of missing test scores is based on the number of tests that should have been carried out in a given year based on the number of children enrolled into relevant grade level mainstream classrooms of the Danish public schools according to the Danish pupil registry. The sample of test observations includes all results of the mandatory and sick test observations with valid PNR.

The sheer size of the schools may influence the implementation of a national testing program. For example, the school size literature often argues that new initiatives are more easily implemented with more homogenous teacher and student bodies (see e.g. Leithwood and Jantzi 2009). The unweighted correlation matrix between school participation rate and school size reveals a slight negative correlation of -0.09 in 2010 (significant at the 1% level) to insignificant -0.04 in 2012. This becomes larger numerically and more significant the larger the participation rate. E.g. for participation rates above 80% the correlation between participation rates and school sizes is -.24 in 2010, -.09 in 2011 and -.13 in 2012 (all significant on the 1% significance level).

Exemptions

In general, pupils within the same class are tested in the same test sessions, including rebooked pupils. Dispensation of pupils may be granted if the school, in agreement with the parents, believes that a pupil is unable to obtain a result that is useful in the evaluation of the child's teaching plan. Of course, pupils exempt from teaching in the specific subject are also excused. Thus, pupils from the lower part of the skill distribution would typically be granted dispensation. Information on pupils who have been exempted from

test taking is available for 2010-2012. This is summarized in Table 4.4. Note that pupils in segregated special education classrooms or schools account for the majority of exemptions granted.

The size of these numbers does not affect the missing test score fraction from above particularly. The fraction of missing test results of Table 4.3 that is explained by pupils being exempt is 7% in 2010, 5% in 2011 and 12% in 2012. Thus, a large fraction of the missing results is likely caused by disobedience of pupils (or teachers). Legitimate exemptions seem to have become more popular with time as almost 2.5 times more dispensations have been granted in 2012 compared to 2010. At the same time, they explain a larger fraction of the missing test scores. Exemptions are typically given more frequently in the higher grades.

Table 4.4 Legitimate exemptions distributed by school and classroom characteristics

School and class type \ year	2010	2011	2012
Regular public school, mainstream classroom	598	1,255	2,493
Regular public school, mixed grade classroom	194	268	748
Regular public school, special education classroom	3,743	4,657	7,317
Segregated special education school	1,910	2,964	4,089
Missing class affiliation	187	377	1,009
Total number of exemptions	6,632	9,521	15,656

Notes. The table summarizes the number of test observations that are legitimately missing from the test data. The exemption data have been matched with school and classroom affiliation from the Danish pupil registry. Data on exemptions are only available until 2012.

Sick tests

The national tests are mandatory and completed in January through April with a retesting period for absentees in June (see section 2.1 for an overview of exact test periods). Table 4.5 shows the number of pupils tested in the ordinary and sick periods. Apart from 2013 only around 4 percent of the test observations are obtained in the sick periods each year. Furthermore, pupils tested outside of the planned test periods are all tested on the same date[24], suggesting that the mandatory test period may have been slightly extended in these years.

Table 4.5 Tests carried out in the ordinary and sick test periods

Period \ year	2010	2011	2012	2013
Ordinary (mandatory) period	441,082	487,113	483,972	228,274
Sick period	21,681	16,524	19,480	261,103
Outside periods	-	113	18	-
Total	462,763	503,750	503,470	489,377

Notes. The sample includes all mandatory and sick test observations from mainstream class room pupils with valid PNR values in 2010-2013. See footnote 6 for an overview of predefined test dates.

[24] January 7 in 2011 and January 9 in 2012.

Two week technical break down in 2010

In 2010 the IT-system suffered from a break down during March 2 – 10, implying that approximately half of the pupils that should have been tested in this period, were not (Rambøll 2013). Further, the 2013 evaluation report states that the test system was highly unreliable on March 1 and 2, and that the tests carried out on March 11 and 12 were made voluntary. The report further notes that 21,697 pupils, who should have been tested in reading, were directly affected by the breakdown (i.e. in the period March 2-9). Approximately half of these were subsequently re-tested. Another 10,286 pupils were booked for reading tests on March 1-2 and 11-12. After the break down affected teachers were encouraged to re-book new test sessions, but on a voluntary basis. Based on descriptive characteristics and regression analyses the 2013 evaluation committee finds that being subject to the technical breakdown was random (Rambøll 2013).

Lockout of teachers in 2013

The 2013 lockout of teachers was a nationwide conflict affecting roughly 80% of the teaching staff in the Danish public schools. It ran 26 days from April 1-26 and obstructed much of the national test program meanwhile. As a response the sick period was prolonged by 2 weeks and schools were required to rebook a test session for pupils affected by the lockout (Kvalitets- og Tilsynsstyrelsen, 2013). As illustrated in Table 4.5 more than half of the tests were carried out in the prolonged re-testing period.

5. Empirical analysis

In this section we first investigate who participates in the tests and how the test results of pupils are associated with background and parental characteristics such as education, income and immigrant status. Then, we investigate how well previous test results predict test scores later on as well as the 9th grade examination marks. We emphasize that this analysis is purely descriptive and are not claiming any causal relationship. The analysis is meant as a preliminary analysis to document some basic characteristics of the pupil proficiency in public compulsory school which have been made possible by the national test program[25].

To enable us to consider overall pupil proficiency within subjects across cohorts and tests, we construct a standardized test score that incorporates measures of pupil skills within all profile areas. Often, information about pupil proficiency within particular profile areas may be too specific compared to our interests.

First, skill level estimates are standardized to mean zero and standard deviation one within each profile area, test and year. These are then averaged within each test for the pupils, and this mean is once again standardized by test and year (see Appendix C). We use this measure as the estimated pupil proficiency within a given test throughout the rest of the paper. By standardizing skill level estimates before averaging

[25] All characteristics are matched from administrative registers maintained by Statistics Denmark. Information of family structure and parental characteristics are measured in the year the child turned five, i.e. before starting school. Individual information on diagnoses and special education needs is from the previous school year. Absence information is measured in August-December of the same school year, i.e. before test taking. See Appendix D for a complete list of covariates.

across profile areas within tests, one ensures that the skill measures within the profile areas are comparable. In practice the distributions of raw test results are very different even for profile areas in the same test (Appendix B). Throughout the literature within economics of education, effects on academic achievement are generally measured in standard deviation-metrics. By standardizing skill level estimates to mean zero and unit standard deviation within profile area, test, and year, coefficients can be readily interpreted as standard deviations. Among others, Lee and Smith (1997) use test scores from adaptive tests based on item response theory[26]. Likewise, they standardize the test results across cohorts and grades and report their results in standard deviation-metric. But there are also many examples of reporting results based on test scores from non-adaptive tests (ex. IQ-test or SATs) in standard deviations (e.g. Fredriksson et al., 2013, Bloom et al., 2008).

Pupil proficiency as measured by the national tests is moderately correlated across profile areas within the same test (between 0.55 and 0.81). Given that the tests are designed to measure different dimensions of skills for each profile area separately, it is clear that there exist a significant relationship between the measured set of skills within profile areas. This is potentially caused by underlying attributes such as motivation. Note that, in accordance with the Rasch requirements, controlling for pupil background characteristics only slightly reduces this correlation. Further, the correlation coefficients seldom reach 0.75 or above, thus, suggesting that the skills measured in the separate profile areas are not complete overlapping and may in fact compose different skill dimensions. The average standardized skill measure, we construct within each test (henceforth referred to as the pupil's test result), is typically correlated with the separate profile area skills of a magnitude of 0.82-0.91. Correlation matrices are presented in Appendix E.

5.1 Test participation

Generally, one would expect that certain pupil characteristics may be correlated with the likelihood of obtaining a test result in the mandatory national tests. We will consider this compromise of a non-selective sample before addressing the other empirical issues.

Table 5.1 presents the results of separate logit regressions of the probability of being exempt from test taking (column 1) and having an 'illegitimate' missing test result (column 2) in any of the national tests, on pupil characteristics. Coefficients denote average partial effects and are shown for a selected subset of covariates (see Appendix DE for a complete list of covariates and sample means). Overall, results are similar across the two outcomes, but as the number of exemptions is very low, the corresponding partial effects are generally very small although still statistically significant. Only girls are significantly less likely to obtain exemptions while they are more likely to be absent from the tests.

[26] See footnote 17 and section 3.1.

Table 5.1 Average partial effects (logit), likelihood of being exempt or having missed a mandatory test

	(1)		(2)	
Selected covariates	Exempt		Illegitimately missing	
Girl	-0.0003	**	0.0019	***
	(0.000)		(0.000)	
Western immigrants/decendants	0.0006	*	0.0087	***
	(0.000)		(0.002)	
Non-Western immigrants/descendants	-0.0003	*	-0.0085	***
	(0.000)		(0.001)	
Low birthweight (<2500 g)	0.0006	**	0.0026	**
	(0.000)		(0.001)	
No. of siblings	-0.0001	*	-0.0023	***
	(0.000)		(0.000)	
ADHD diagnosed	0.0010	**	0.0108	***
	(0.000)		(0.004)	
Special education needs, primary cause				
Learning disability	0.0038	***	0.0335	***
	(0.000)		(0.001)	
Mental disability	0.0043	***	0.0485	***
	(0.000)		(0.004)	
Social disability	0.0034	***	0.0381	***
	(0.001)		(0.006)	
Physical disability	0.0059	***	0.0505	***
	(0.001)		(0.007)	
Other	0.0030	***	0.0320	***
	(0.000)		(0.001)	
Family information				
Single mom	0.0005	***	0.0154	***
	(0.000)		(0.001)	
Mother's age	0.0000		-0.0005	***
	(0.000)		(0.000)	
Father's education:				
≤ high school	-0.0000		-0.0021	
	(0.000)		(0.002)	
Vocational	-0.0009	**	-0.0111	***
	(0.000)		(0.002)	
Bachelor	-0.0010	**	-0.0138	***
	(0.000)		(0.002)	
Higher	-0.0014	***	-0.0093	***
	(0.001)		(0.002)	
Father's logearnings	-0.0003	***	-0.0026	***
	(0.000)		(0.000)	
Capital area school	-0.0001		0.0319	***
	(0.000)		(0.001)	
Constant	-0.0108	***	-0.0184	***
	(0.003)		(0.002)	
N	1,600,238		2,116,150	
Mean outcome	0.0024		0.0740	
Pseudo R-squared	0.127		0.105	

Notes. Results in (1) are based on the test years of 2010-2012, while (2) is based on the full sample. In addition to the control variables listed in the table, both specification include year and test fixed effect and the remaining controls from Table D.1 (except absence information). Standard errors clustered on individuals are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

By construction of the exemption criteria, weak pupils are more likely to be exempted from the test e.g. because of insufficient proficiencies in Danish or exemption from the relevant subject. Not surprisingly, pupils with special education needs in the previous year have a significantly higher probability of being exempt. Being of non-Western background does not increase the likelihood of being exempted from the national tests once family and child characteristics are controlled for. Further, we find that it decreases the probability of missing a test. Otherwise, particularly having a father who obtained more than a high school degree significantly reduces the likelihood of not being tested across outcomes.

5.2 Pupil Skill Level Explained

Table 5.2 presents the selected results of regressing pupil test results on pupil background characteristics. Across all subjects observed pupil and parental background explain around 13 to 21% of the variability in the pupil proficiency as measured by the national tests. This is of the same magnitude as other associations between pupil test scores and parental background (see e.g. Schochet 2008). Based on the results from Table 5.2, higher reading scores are generally associated with being a girl; however, the advantage is decreasing over grades. Also the pattern is reversed for English as well as the science-based subjects; in particular boys seem to excel in physics/chemistry. Non-Western immigrants or descendants hereof are generally of significantly lower proficiency in all tested subjects, though to a lesser extent in English.

Another well-established pattern emerges; lower proficiency as measured by the tests is associated with low birth weight, being assigned to special needs education and lower socioeconomic status as represented by parents' log earnings and education (see e.g. Hanushek and Woessmann 2011; Carneiro and Heckman 2003; Björklund and Salvanes 2011). Results are unchanged when pupil absence information is included.

The four years of national test data allow us to compare test results of the same individuals within Danish, reading and math across two grade levels. In Table 5.3 we investigate how well pupil achievement in later grades is explained by previous achievement on the tests. For each test result, e.g. reading in grade 4 (column 1), we run two specifications: In specification a) we only include the pupils previously measured test result of the same subject (here, measured in grade 2), and in specification b) we add all characteristics from Appendix E except absence information. See Table 2.1 for an overview of the previous tests.

The magnitudes of the coefficients indicate very high correlations between previous and current test results. Generally, increasing previous test result by 1 standard deviation (SD) improves test results by 0.6-0.7 SD (columns a). Interestingly, this relation is as good as unchanged when pupil and family characteristics are included (columns b). Also the R-squared values are relatively stable; previous reading score explains more than 50% of the variability in the current test results, though a little less for the math test. Adding a wide range of controls to the specification increases the R-squared with just around 1.5 percentage points (4.5 for math). Thus, parental and pupil characteristics add very little explanatory power to our model. In general, the

coefficients on pupil and parental characteristics of Table 5.3 (omitted) are smaller and less significant compared to Table 5.2.

Table 5.2 OLS estimates, test results explained by parental and pupil characteristics (linguistic tests)

Selected covariates	(1) Reading, grade 2	(2) Reading, grade 4	(3) Reading, grade 6	(4) Reading, grade 8	(5) English, 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)	(0.024)	(0.020)	(0.026)	(0.018)
Family information					
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)
Father's logearnings	0.025 ***	0.022 ***	0.018 ***	0.019 ***	0.028 ***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Capital area school	-0.088 ***	-0.021	-0.008	-0.058 **	-0.132 ***
	(0.024)	(0.028)	(0.029)	(0.028)	(0.026)
N	199,991	202,823	204,894	188,819	195,185
Mean outcome	0.017	0.032	0.039	0.052	0.025
Adjusted R-squared	0.148	0.181	0.206	0.199	0.168

Notes, see Table 5.2 (continued)

Table 5.2 (continued) OLS estimates, test results explained by parental and pupil characteristics (science tests)

Selected 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 immigrant/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)
N	203,578	204,810	186,631	186,170	186,346
Mean outcome	0.017	0.034	0.020	0.024	0.026
Adjusted R-squared	0.125	0.162	0.135	0.154	0.172

Notes. In addition to the control variables listed in the table, all specification include year fixed effect and the remaining controls from Table D.1 (except absence information). Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

As family structure and birth information are measured before the pupil enters compulsory schooling, these are in some sense already incorporated in the coefficients on previous test result. However, it is worth noting

that previous test result alone explains more than 34 percentage points more of the variability in the reading scores compared to all the other skill determinants combined. As the tasks are similar across tests, though the item bank changes, this presumably explains some of this difference in explanatory power. Note that parental background is undoubtedly highly related to innate pupil ability/ability when entering compulsory school (Table 5.2, the earliest measures of pupil proficiency is grade 2 and grade 3 for reading and math, respectively). Thus, research on the determinants of skill level in the years before compulsory school is also highly relevant in the Danish setting. For a discussion see e.g. Heckman (2006).

Table 5.3 OLS estimates, test results explained by previous test result in same subject

	(1a)	(1b)	(2a)	(2b)
	Reading, grade 4	Reading, grade 4	Reading, grade 6	Reading, grade 6
Previous test result	0.686 ***	0.621 ***	0.760 ***	0.703 ***
	(0.004)	(0.004)	(0.004)	(0.004)
Constant	0.028 ***	-0.451 ***	0.031 ***	-0.411 ***
	(0.006)	(0.122)	(0.006)	(0.064)
N	90,194	90,194	92,922	92,922
Mean outcome	0.058	0.058	0.067	0.067
Adjusted R-squared	0.491	0.515	0.589	0.605
Controls	NO	YES	NO	YES
	(3a)	(3b)	(4a)	(4b)
	Reading, grade 8	Reading, grade 8	Math, grade 6	Math, grade 6
Previous test result	0.744 ***	0.690 ***	0.592 ***	0.517 ***
	(0.004)	(0.004)	(0.007)	(0.006)
Constant	0.026 ***	-0.780 ***	0.045 ***	-0.549 ***
	(0.006)	(0.089)	(0.009)	(0.116)
N	87,110	87,110	43,827	43,827
Mean outcome	0.089	0.089	0.067	0.067
Adjusted R-squared	0.573	0.589	0.358	0.403
Controls	NO	YES	NO	YES

Notes. Results are conditional on having obtained a previous test result. In addition to the control variables listed in the table, all specification include year fixed effect and the remaining controls from Table D.1 (except absence information). For each column, "previous test result" denotes grade 2 reading results when the outcome variable is the grade 4 reading result, grade 4 reading result when the outcome is grade 6 reading result etc. Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table 5.3 indicates that a large group of low-skilled pupils, irrespectively of family background, are 'stuck' in the lower end of the test score distribution across years. Thus, having more or less advantageous family background does not seem to explain much of the progress from, for example, grade 2 to grade 4 reading results. In other words, only to a lesser extent will the reading scores of high SES children improve compared to those of other children with the same previous skills and vice versa. To exemplify, pupils' ranks within the test score distributions are illustrated in Table 5.4. Here, pupils are divided into three overall groups: the bottom 25%, the middle 50% and the top 25%. The diagonal then illustrates the number of pupils who are 'stuck' in the test score distribution. It is important to keep in mind that, because of the standardized

test results, this is *relative* to other pupils. Thus, pupils may very well have progressed since the previous test but not relatively more than other pupils. As such, more than 65% of the pupils in each group are found in the same group two years after. Approximately 30% transfer to an adjacent group whereas only 2% switch from the lower to the upper quartile from grade 6 to 8. Also when dividing into smaller intervals the pattern is very clear.

Table 5.4 demonstrates the rank transitions between the 6th and 8th grade reading tests alone, but as suggested by Table 5.3 other matrices are quite similar. As this is only an outset for future research, the question of who manages to transfer to higher test score groups still remains. However, the national biannual reading tests greatly improve the possibility to uncover such patterns.

Table 5.4 The development from grade 6 to grade 8 reading test results

Percentile, grade 6	Percentile, grade 8 reading			Total
	≤ 25	26-75	> 75	
≤ 25	13,763	6,212	395	20,370
	68%	30%	2%	100%
26-75	6,782	30,253	7,068	44,103
	15%	69%	16%	100%
> 75	263	7,477	14,877	22,617
	1%	33%	66%	100%
Total	20,808	43,942	22,340	87,090
	24%	50%	26%	100%

Notes. The table includes pupils who have taken both a grade 6 and a grade 8 reading test two years apart.

5.3 What are the national tests measuring?

The national tests are thought to measure true skill level within specific cognitive areas through primary and lower secondary school. But given their relatively young age little evidence exists of the relation between test results and other measures of later success. Overall, results presented in Section 5.2 are very similar to patterns from other standardized tests: girls are slightly better at reading, boys at math-related subjects, high SES children overall etc.

As a preliminary external validity check, we present evidence of the associations between test results and pupils' 9th grade exit examination marks. Exit examination results are generally considered to be highly associated with some adult success measures, e.g. successfully finishing vocational college (Hvidtfeldt and Tranæs 2013). Information of exit examination marks is available until 2013, thus, we are able to link 9th grade examination marks to individuals' test results from grades 6 through 8 for up to three cohorts of pupils.

The 9th grade exit exams consist of mandatory examinations in the subjects Danish (reading, writing, spelling and oral performance), math (calculus and problem solving), English (oral) and physics/chemistry (oral). Subjects, such as geography, biology, math (oral), English (written) etc., are decided by a random draw and

are omitted from this analysis. Here, we construct average measures of pupils' examination marks in Danish and math[27].

Exit examination marks and test results may differ for various reasons. First of all, pupil skill level is objectively estimated on a continuous scale while examinations marks are graded by a teacher and a censor on a 7-point ordinal scale ranging from -03 to 12. Secondly, in case of the oral exams pupils typically draw only one or two topics from the curriculum to present. Compared to this, the national test contains items that are relevant for the specific cognitive profile area on a more general scale (the questions compose a series of random draws of single items within profile areas). Further, as the purpose of the exit exams and the national test differ, they are likely to measure somewhat different sets of skills. Finally, there is the issue of high versus low stake testing: The test environment of the national tests is considerably more informal compared to that of the exit exams. Pupils are allowed to take a break, interact (to some degree) with their teacher, and the tests are typically carried out in a regular classroom equipped with computers and in the presence of the classmates. Compared to this, entire cohorts are usually put together when taking their 9th grade written exams – perhaps in facilities outside of school. Also, the stakes of the national tests are quite low seeing as the primary purpose of the test results is to evaluate the teaching needs of the pupil. This may cause some pupils to perform better as exam jitters are less pronounced, while others may perform poorer because stakes are low. Also at the school or even class level, the national tests are of relatively low stake. Neither municipalities nor principals are allowed to rank or sanction schools and teachers based on the national tests.

5.3.1 The national tests and exit exam marks

In Table 5.5 we regress the pupils' 9th grade examination mark on the same-subject national test result obtained in earlier grades. On the individual level previous test results from the national tests explain 48-51% of the variation in Danish and math examination marks. The corresponding proportions of English and physics are 42% and 23%, respectively. In all cases increasing the test result by 1 SD is associated with a 2 grade point increase when excluding other covariates from the model. Thus, the national tests indeed measure some skills that are at least very highly correlated with the 9th grade exit exams[28].

Table 5.6 presents the results of regressing average exit exam marks on relevant test results as well as pupil characteristics. Controlling for baseline covariates reduces the point estimates of the previous test results by approximately 0.2 grades each to around 1.8 while the R-squared rises slightly. Note that some observations are lost, as not all pupils have completed the mandatory exit exams. This is more frequently seen in subjects such as physics/chemistry and foreign languages. Pupils who have missed one or more exams in Danish or math are still present in the sample. Their GPAs are adjusted accordingly.

[27] Pupils are also graded based on the appearance of their written performances (omitted here).

[28] The average standardized test result within subject is more highly correlated with average exit examination marks compared to test results within profile areas alone.

Interestingly, the point estimate to the 6th and 8th grade reading test results are practically identical. Thus, reading scores in grade 6 seems to be just as good a predictor of Danish exit exam achievement as grade 8 reading scores, even though two more years of learning have taken place in between. Further it is worth noting that once baseline characteristics and previous reading proficiency are controlled for, non-Western pupils actually earn higher examination marks compared to native Danes in the linguistic subjects. To dig deeper into this issue, we split the GPA outcome in Danish into separate outcomes for the oral and written performances. We then include an interaction term between 8th grade reading results and an indicator for being non-Western to specification (2) of Table 5.6. Results are shown in Table 5.7. Here, our model explains 28% of the variability in the oral examination results while explaining up to twice as much in written exams, which could be expected given the nature of the tests. The coefficient on the interaction term reveals that the correlation between 8th grade reading scores and both oral and written examination performance is significantly smaller for pupils of non-Western background. Thus, either the 8th grade test results or the 9th grade exit examination marks are a less precise measure of 9th grade proficiency for this group of pupils. The point estimate of being of non-Western background is now significantly negative for the written outcomes reading and essay while significantly positive and equally large in magnitude for the oral and the spelling performances.

Firstly, it should be noted that the oral essay examination marks are based on the performance assessments of a teacher and an outside censor alone. The reading and spelling marks, on the other hand, are subject to much less discretion. The questions are largely either multiple choice or 'fill in the blanks', where the final examination mark is a step function of number of points collected during the exam.

Table 5.5 OLS estimates, 9th grade examination marks on test results, no baseline covariates

	(1)	(2)	(3)	(4)	(5)
	GPA, Danish	GPA, Danish	GPA, math	Exit exam, English	Exit exam, physics
Test results:					
Reading, grade 6	2.010 ***				
	(0.013)				
Reading, grade 8		1.998 ***			
		(0.009)			
Math, grade 6			2.229 ***		
			(0.018)		
English, grade 7				2.421 ***	
				(0.015)	
Physics/chemistry, grade 8					1.831 ***
					(0.021)
Constant	6.575 ***	6.525 ***	6.506 ***	7.212 ***	6.149 ***
	(0.016)	(0.012)	(0.020)	(0.017)	(0.020)
N	46,728	138,970	46,484	91,331	134,929
Mean outcome	6.718	6.663	6.659	7.388	6.236
Adjusted R-squared	0.497	0.507	0.483	0.420	0.227
Covariates	No	No	No	No	No

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

Table 5.6 OLS estimates, 9th grade examination marks on test results and baseline covariates

	(1) GPA, Danish	(2) GPA, Danish	(3) GPA, math	(4) Exit exam, English	(5) Exit exam, physics
Test results:					
Reading, grade 6	1.763 ***				
	(0.014)				
Reading, grade 8		1.762 ***			
		(0.010)			
Math, grade 6			1.960 ***		
			(0.018)		
English, grade 7				2.251 ***	
				(0.016)	
Physics/chemistry, grade 8					1.645 ***
					(0.022)
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)	(0.242)	(0.232)	(0.233)
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)
Constant	0.968 **	2.334 ***	1.997 ***	2.463 ***	1.029 ***
	(0.477)	(0.181)	(0.299)	(0.563)	(0.324)
N	46,728	138,970	46,484	91,331	134,929
Mean outcome	6.718	6.663	6.659	7.388	6.236
Adjusted R-squared	0.582	0.581	0.532	0.445	0.275

Notes. In addition to the control variables listed in the table, all specifications include year fixed effect and the remaining controls from Table D.1 (except absence information). Standard errors clustered on schools are in parentheses. Asterisks indicate statistical significance at the ***1%, **5%, and *10% level, respectively.

Table 5.7 OLS estimates, 9th grade examination marks in Danish on test results and baseline covariates

	(1)	(2)	(3)	(4)
	Oral exam	Written exams		
	Oral	Essay	Spelling	Reading
Test results:				
Reading, grade 8	1.517 ***	1.642 ***	2.067 ***	1.995 ***
	(0.012)	(0.011)	(0.012)	(0.013)
Reading, grade 8 x non-Western immigrant	-0.082 **	-0.190 ***	-0.321 ***	-0.369 ***
	(0.034)	(0.026)	(0.029)	(0.026)
Non-Western immigrant/descendant	0.389 ***	-0.089 **	0.154 ***	-0.380 ***
	(0.051)	(0.041)	(0.040)	(0.036)
Constant	2.231 ***	2.303 ***	2.205 ***	2.449 ***
	(0.314)	(0.237)	(0.225)	(0.221)
N	137,697	138,150	137,986	138,200
Mean outcome	7.456	6.383	6.454	6.416
Adjusted R-squared	0.278	0.395	0.502	0.444
Covariates	YES	YES	YES	YES

Notes. Table note 5.6 applies.

Evidence from the teacher bias literature suggests that certain groups of pupils may be given preferential treatment when teachers in their sole discretion grade pupil performances. For example, Jensen and Smith (2007) suggest that the formation of teacher expectations in schools with a high share of non-Western pupils may lead to systematic differences between non-Western and Danish pupils when comparing 2005 PISA Ethnic scores to yearly marks. Although we should be careful giving the estimates any causal interpretations, column 1 of Table 5.7 indeed reveals that being of non-Western background is associated with better oral examination marks when controlling for 8th grade reading scores. Notice, though, that being non-Western also significantly affects the results of the low discretion exams in columns 3 and 4. This suggests that the test results rather than the exit examination marks that may be imprecise for non-Westerners. Potential explanations include, for example, accumulation of oral and written skills at different paces from grade 8 to grade 9 compared to Danish pupils or failure of objectivity of the test items.

5.4 Test score gaps throughout compulsory school

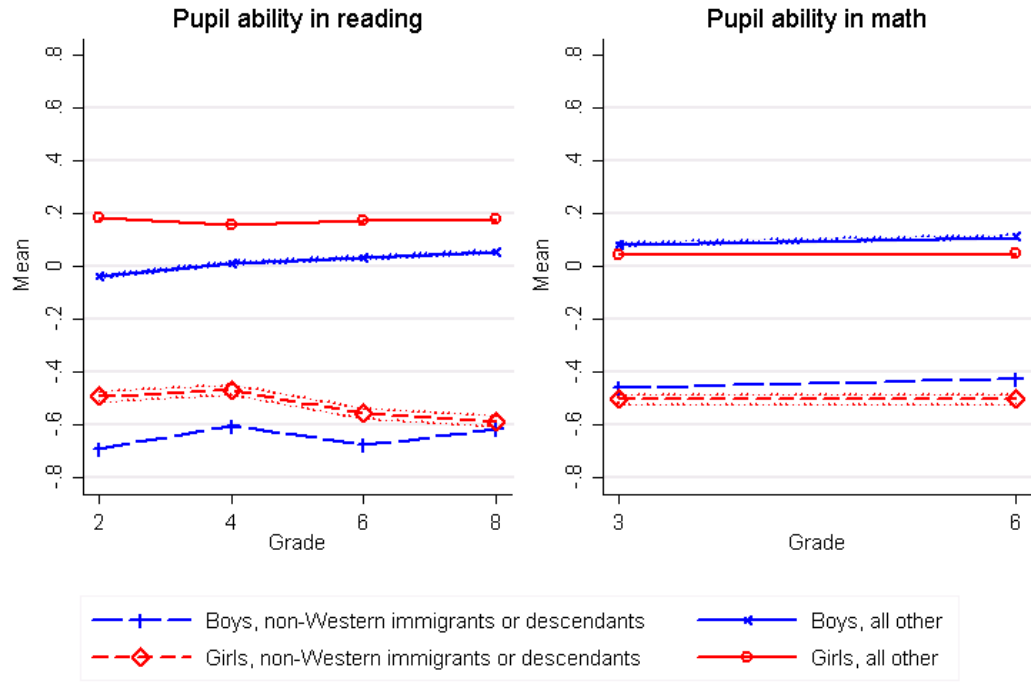
This section presents evidence of pupil proficiency throughout compulsory school, where the focus is on the reading and math tests. From Section 5.2 we learned that there are indications that certain groups of pupils are stuck in the lower end of the skill distribution. In this section we will elaborate further on this but first, recall that test results have been standardized to mean 0 and standard deviation 1 within each test, which means that we can only observe groups of pupils relative to other groups. Thus, if a group of pupils improve over another group, the latter must necessarily worsen relative to the first[29].

[29] In figures of Section 5.4 we have only included pupils in regular public schools. Thus, the mean of the standardized test results are slightly greater than zero in all subjects.

The left (right) panel of Figure 5.1 illustrates the average reading (math) scores of boys and girls divided by ethnic background. In line with the findings of Table 5.2, girls generally perform better than boys in reading. Although for pupils of non-Western background the difference in means disappear across grade levels. Further, between grades 2 and 4 the average reading scores of boys and girls seem to converge slightly. On the other hand, the average differences in math scores are fairly constant across grades. Overall, the test results of pupils from non-Western countries are considerably lower than others' without any sign of convergence.

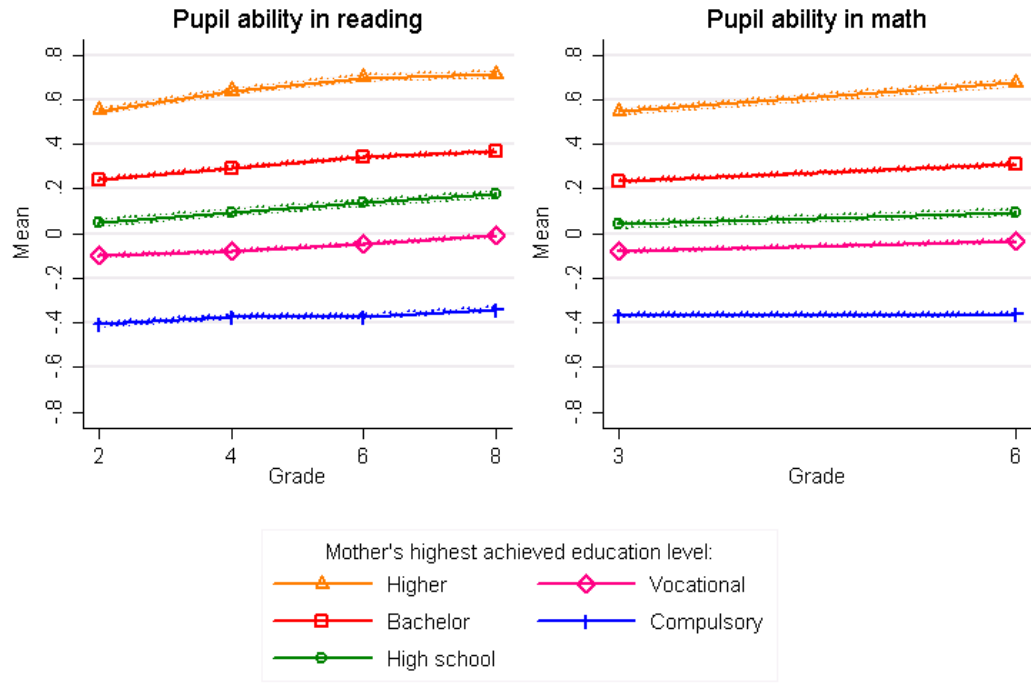
If we segregate pupils after their socioeconomic status as proxied by parental education and income, the pattern is equally clear. Again the left (right) panels of Figure 5.2 and Figure 5.3 depict average reading (math) scores. Both when pupils are divided by parental education and parents' highest earnings the gaps in average test results within the groups are very stable across grade levels. Test results in the 2nd grade is clearly ranked by socioeconomic status, and, consistent with other findings presented in this paper, relative to other groups of pupils the advantage of background characteristics persist. Pupils are more or less 'stuck' in the skill distribution.

Figure 5.1 Average test results in reading and math by grade and gender and immigration background



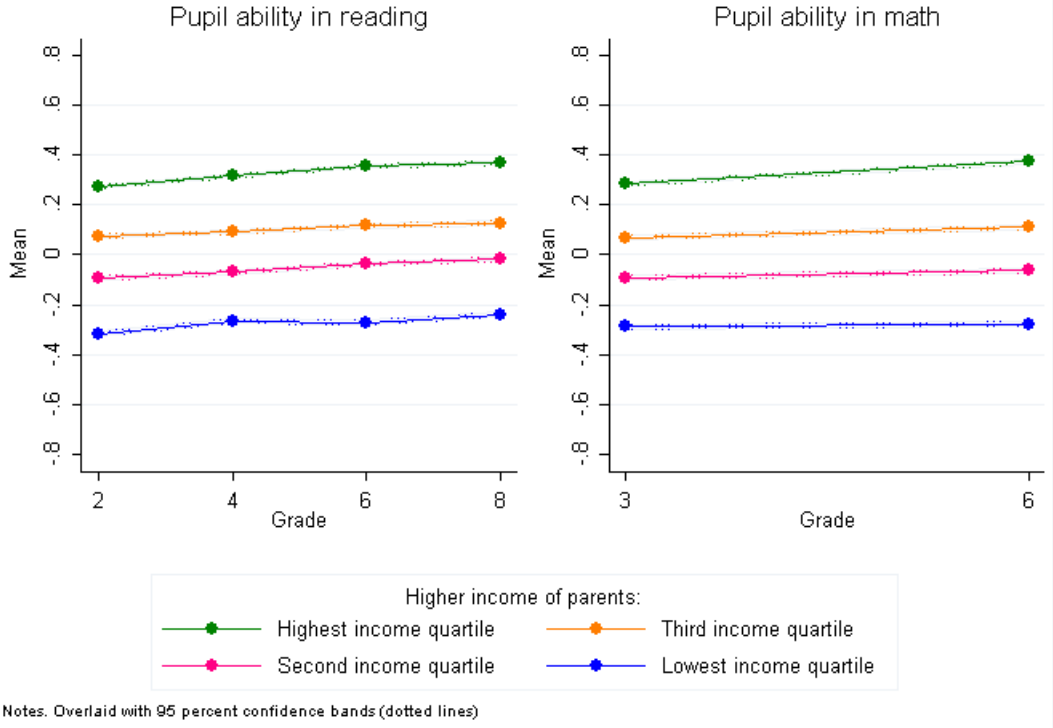
Notes. Overlaid with 95 percent confidence bands (dotted lines)

Figure 5.2 Average test results in reading and math by grade and mother's highest education level



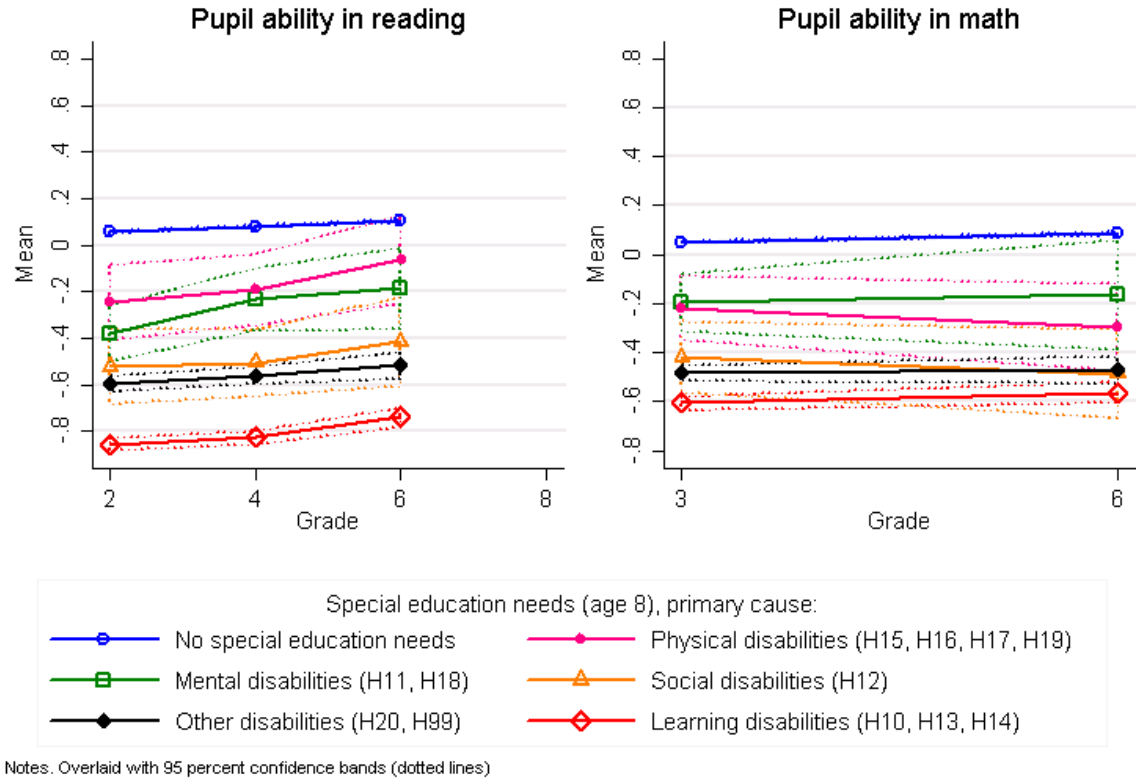
Notes. Overlaid with 95 percent confidence bands (dotted lines)

Figure 5.3 Average test results in reading and math by grade and parent's income level



Notes. Overlaid with 95 percent confidence bands (dotted lines)

Figure 5.4 Average test results in reading and math by grade and special education needs



Notes. 8th grade observations are omitted due to lack of special education data for this age group. Disability categories are aggregates of the 12 official disability categories; H10: Learning disorder, H11: Mental and behavioral disorders, H12: Socio-emotional disorders, H13: Reading/writing disability, H14: Speech/language disability, H15: Hearing impairment, H16: Vision impairment, H17: Physical disability, H18: Psychiatric disorders, H20: Other disability, and H99: Not specified.

Figure 5.1 to 5.3 resemble figures from the international literature on child development and educational achievement. 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.

Additionally, we have access to individual data on special education needs: primary cause and hours of special education training inside or parallel with the regular classroom. The last figure, Figure 5.4, illustrates the average test results across grades dividing pupils into five categories based on their type of special education needs: Physical disabilities, mental disabilities, social disabilities, learning disabilities and other causes. All referrals are measured at the age 8, i.e. before the earliest national tests are carried out (grade 2)[30]. For comparison, the blue line represents the average test result across grades for students with no documented special education needs at age 8. In grade 2 the average test score gap is smallest between non-

[30] This definition differs from the previous year assignment to special needs education-covariates presented in Table E.1 and used in the regressions. 5.0% of the 2nd grade sample, 4.8% of the 4th grade sample and 3.7% of the 6th grade sample have perceived special education needs at age 8. Conditional on being referred to special needs education, the primary causes are distributed as follows: 52% has learning disabilities, 40% other/unspecified disabilities, 4% mental disabilities, 2% social disabilities and 2% physical disabilities. Some pupils with special education needs may not be officially classified already at age 8. A pedagogical and psychological team based on a request by either the school principal or parents assesses special education needs. Disability categories are aggregates of the 12 official disability categories, see figure note.

referred pupils and pupils referred because of physical disabilities (approximately 0.2 SD, insignificant in grade 6). The gap is a little larger for pupils referred because of mental disabilities while it is roughly the same for referrals because of social and other/unspecified disabilities. The very largest gap is found for pupils who are referred because of learning disabilities; they score almost one standard deviation below their peers with no special education needs at age 8. In a recent paper, Feng and Sass (2013) show that pupils who are taught in a mix of mainstream and special education classes (comparable to our special needs education pupils) on average score 0.9 SD below their peers in math and reading, respectively. Overall, the test score gaps in Figure 5.4 decrease by 0.1-0.2 SD across cohorts. However, this convergence may in part be driven by some of the pupils with the most detrimental disabilities transferring to segregated special education classes or schools across grades.

Figure 5.4 suggests that pupils with documented disabilities at age 8, despite being assigned to special needs education, only to a small degree are catching up with their average peers over the years. However equally important is the fact that the test score gap does not widen. Compared with the American literature, only a few US states have individual level data on pupils with special education needs. Using Texan data, Hanushek and Rivkin (2002) show that the reading score gap between regular pupils and pupils who are emotionally disabled widens from 0.69 SD in grade 4 to 0.95 SD in grade 7, while the corresponding regular/learning disability test score gap in math widens from 0.84 in grade 4 to 1.07 standard deviations in grade 7.

6. Concluding remarks

The value of the national tests as seen from an empirical researcher cannot be denied. Particularly, the biannually repeated reading tests from grades 2 through 8 offer great possibilities and as such the test results have already been applied in multiple working papers. With this paper we share our insights and experiences from working with the first four rounds of national test data as a starting point for the discussion. We have explained in detail the test process and summarized the main strengths and limitations of the tests. For future assessment and evaluations of empirical analysis, we find that some practical guidelines on how to interpret and use the test results will be valuable.

This paper uncovers how to interpret the estimated skill level estimates as well as general data issues to consider when working with the test data. Apart from the likely consequences of the nation-wide teacher lock out in April 2013, we find that the test participation is increasing across years. For second graders in 2012, only 2.18% of the sample did not obtain a test score. Overall, the participation rates are declining across grades. Further, evidence suggests that missing the tests is generally correlated with low socio-economic status.

Depending on the question of interest, different transformations of the raw test results are relevant. The transformed points from 1 to 100 are generally relevant as a policy device to follow the achievement of e.g. second graders over time, because the bounds on the points are constant across years. However, for research purposes, where the interest lies in following or estimating the progress of individuals (compared to his or her peers), using the more precise skill measures is often preferable.

We have provided the reader with preliminary analyses of the relation between pupils' national test results, pupil background characteristics and 9th grade exam results. We find the expected relationship between pupil background and performance on the national tests: girls outperform boys in reading while boys are slightly better at math-related subjects. Further, lower test results are associated with low birth weight, special education needs and lower socioeconomic status as represented by parents' logearnings and education. Across all subjects observed pupil and parental background explains around 13 to 21% of the variability in the test results. Using pupils' national test results in earlier grades we are now able to explain more than 50% of the variability in later test performance as well as 9th grade examination result. This feature is quite remarkable.

Our analyses show that the national tests are able to measure skills that are at least very highly correlated with the skills measured by the 9th grade examination marks. Generally, a one standard deviation increase in pupil achievement on the national tests is associated with a two grade point increase in the 9th grade exam marks. We interpret these findings as a preliminary validity test of the national tests as a valid predictor for future performance in the same way as the 9th grade exit examination marks. The next important step will be to show if the change in test results is an equally good predictor of a change in exit examination marks.

Finally, consecutive testing of pupils in the same subjects enables us to study the test score gaps across grade levels. Transitional analyses suggest that more than two thirds of the Danish pupils are 'trapped' in the same part of the skill distribution compared to two years earlier. Further, we document a similar but worrying stability in average test score gaps between pupils of different socio-economic background as in the American literature. Additionally, the graphical evidence suggests that pupils with documented special education needs at age 8 largely do not seem to catch up with the average peers throughout compulsory schooling.

Summing up, the introduction of the national test program with its ten mandatory tests evaluating pupil proficiency within six subjects and 18 cognitive areas, allows us to follow and investigate the cognitive skill formation in compulsory schooling more closely. These features are of great importance when considering future investments in the Danish public school system.

7. Literature

Andersen, E. (2002), "Some New and Some Old Results for the Polytomous Rasch Model", *Classification, Automation, and New Media*, Springer-Verlag Berlin.

Beuchert, L.V., Humlum, M. K., Nielsen, H.S., and Smith, N. (2014), "The Short-term Effect of School Consolidation on Student Achievement", Aarhus University, Unpublished manuscript

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

Bloom, H. S., Hill, C. J., Black, A. B., and Lipsey, M. W. (2008). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. *Journal of Research on Educational Effectiveness*, *1*(4): 289-328.

Bond, T. and C. Fox. 2007. "Applying the Rasch Model." Second edition. *Routledge*.

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

Danmarks Evalueringsinstitut (2002), "Evaluering af folkeskolens afgangsprøver – Karakterundersøgelse", EVA - Danmarks Evalueringsinstitut.

Dee, T. (2007), "Teachers and the Gender Gaps in Student Achievement", *The Journal of Human Resources*, Vol. 42, No. 3, pp. 528-554.

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

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

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

Figlio, D. and S. Loeb (2011), "School Accountability." Ch. 8 in E. A. Hanushek, S. Machin and L. Woessmann (eds.), *Handbooks in Economics*, vol. 3., 383-421.

Fredriksson, P., Öckert, B., and Oosterbeek, H., (2013), "Long-term Effects of Class Size". *Quarterly Journal of Economics*, 128(1), 249-285.

Hanushek, E. A, Kain, J.F., 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, Volume 3*

Heckman, J. (2006), ”Skill Formation and the Economics of Investing in Disadvantaged Children”, *Science, Vol 312*

Hvidtfeldt, C. and T. Tranæs (2013), ”Folkeskolekarakterer og Succes på Erhvervsuddannelserne”, Rockwool Fondens Forskningsenhed, Working paper no. 61.

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

Jensen, V. M. and L. P. Nielsen, (2010) ”Veje til Ungdomsuddannelse 1: Statistiske analyser af folkeskolens betydning for unges påbegyndelse og gennemførelse af en ungdomsuddannelse”, SFI - Det Nationale Forskningscenter for velfærd, 10:24

Jensen, P. and N. Smith (2007), ”Etnisk koncentration i folkeskolen”, PISA Etnisk 2005, Egelund, N. and T. Tranæs (Eds), *Rockwool Fondens Forskningsenhed og Syddansk Universitetsforlag*.

Kreiner, S. and J. Wandall (2012), “Misforståelser om de nationale test”, *Folkeskolen.dk* (April 2012), available at http://www.folkeskolen.dk/~/5/5/misforstaaelser-om-de-nationale-test.pdf

Kristoffersen, J. H. G., M. Krægpøth, H. S. Nielsen and M. Simonsen (2015), Disruptive School Peers and Student Outcomes. *Economics of Education Review* 45: 1-13.

Kvalitets- og Tilsynsstyrelsen (2013). ”Information om obligatoriske test efter lockout”, (April 29, 2013) Sags nr.: 026.53P.271. Available at http://www.uvm.dk/I-fokus/OK13/~/media/UVM/Filer/%20fokus/Tema /OK13/130429%20Information %20om%20obligatoriske%20test%20efter%20lockout.ashx

Lee, V. and J. Smith (1997), “High School Size, Which Works Best and for Whom?”, *Educational Evaluation and Policy Analysis*, Vol. 19, pp. 205-277.

Leithwood, K., & Jantzi, D. (2009), “A Review of Empirical Evidence About School Size Effects: A Policy Perspective”, *Review of Educational Research,* s. 464-490.

OECD (2004), “Reviews of National Policies for Education: Denmark – Lessons from PISA 2000”, *OECD Publishing*. Available at http://dx.doi.org/10.1787/9789264017948-en.

Pøhler, L. and S. Sørensen (2010), ”Nationale test og anden evaluering af elevens læsning”, *Dafolo Forlag*, 1st edition.

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%20Evalueri ng%20af%20de%20nationale%20test_rapport.ashx

Rambøll (2014), ”Supplement til evaluering af de nationale test – Rapport”, Available at http://uvm.dk/~/media/UVM/Filer/Udd/Folke/PDF14/Aug/140827%20Supplement%20til%20evaluering%2 0af%20de%20nationale%20test%20Endelig%2018%2008%2014.pdf

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)

Schochet, P. (2008). "Statistical power for random assignment evaluations of education programs". *Journal of Educational and Behavioral Statistics*, 33, 62–87

Skolestyrelsen (2010a), "National test dansk, læsning – 2. 4. 6. og 8. klasse", Styrelsen for Evaluering og Kvalitetsudvikling af grundskolen, Kontor for afgangsprøver, test og evalueringer.

Skolestyrelsen (2010b), "De Nationale test og Kommunen – brug af testresultater i kommunernes kvalitestarbejde", *Styrelsen for Evaluering og Kvalitetsudvikling af grundskolen, Kontor for afgangsprøver, test og evalueringer*.

Skolestyrelsen (2011a), Kvalitetsrapporten som kommunalt styringsredskab (In English: The Quality Assessment as an Accountability Measure), Kontor for Kvalitetssikring og Kvalitetsudvikling, Skolestyrelsen. Link.

Skolestyrelsen (2011c), Brug testresultaterne – inspiration til pædagogisk brug af resultater fra de nationale test, Kontor for Kvalitetssikring og Kvalitetsudvikling, Skolestyrelsen.

Todd, P. and K. Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *The Economic Journal* **113** (Feb): F3-F33.

Undervisningsministeriet (2010), "Fælles Mål 2009 – Elevernes alsidige udvikling", *Undervisningsministeriets håndbogsserie*, nr. 4-2010

Undervisningsministeriet (2012a), "Vejledning til prøverne i faget fysik/kemi", Publication: *Ministeriet for Børn of Undervisning, Kvalitets- og Tilsynsstyrelsen*.

Undervisningsministeriet (2012b), "Orientering om folkeskolens afsluttende prøver 2011/2012", Publication: *Ministeriet for Børn of Undervisning, Kvalitets- og Tilsynsstyrelsen*.

Undervisningsministeriet (2014), "Statistisk usikkerhed i de nationale test". Available at: http://uvm.dk/Aktuelt/~/UVM-DK/Content/News/Udd/Folke/2014/Jan/140127-Statistisk-usikkerhed-i-de-nationale-test.

Undervisningsministeriet (2014b), "Analyse af effekten af elevers social baggrund", Hentet fra: http://www.uvm.dk/~/media/UVM/Filer/Udd/Gym/GYMudspil/141127_Betydning_af_elevernes_sociale_baggrund.pdf

Undervisningsministeriet (2015, January), "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)

UNI-C (Strange, L., C. Jensen, H. Albeck and J. Lund) (2012a), ”Den Adaptive Algoritme i De Nationale Test”, *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)

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

Appendix A. Transformation of raw test results to points

Table A.1 Parameter values of the sigmoid function used in the transformation from skill level estimates to points

Fagid	Profile area	e	f (median)	k	No. of estimates	Max. lower deviation
102	1	1.885385	1.008282157	1.71175	15221	1.838909508
102	2	1.241358	1.48779365	0.828111	15199	4.573852237
102	3	0.884593	0.569144774	1.140831	15189	2.720117826
104	1	1.539668	0.16471296	1.972179	18221	1.594448207
104	2	1.470999	2.172235604	1.370844	18196	2.933474224
104	3	1.333277	0.611017536	1.645153	18173	0.994413886
106	1	1.916657	-0.078132367	1.899646	17919	2.676404106
106	2	1.214495	2.626452613	1.545634	17916	1.56093862
106	3	1.71657	0.755353712	1.684322	17912	1.677220206
108	1	1.942046	0.775059039	2.14648	17682	2.795040376
108	2	1.201171	4.254865643	1.9204	17674	4.740449136
108	3	1.619268	0.941196242	1.703211	17644	1.962496252
203	1	1.972083	0.632407365	1.847906	20142	2.323573072
203	2	2.593047	0.477568751	1.742104	20122	3.874305604
203	3	1.463831	0.835858159	1.604687	20097	1.857471774
206	1	2.412709	0.377331183	1.990316	19719	0.919014229
206	2	2.598276	0.030194617	2.319469	19712	1.68646564
206	3	2.221905	-0.238599876	1.97469	19700	2.06753292
308	1	3.96054	-0.344441938	2.750938	17050	1.597025998
308	2	4.979514	-0.262299546	2.642038	17032	3.342283871
308	3	4.029897	-0.245534994	3.158241	17023	1.464795685
408	1	3.699929	-0.180678965	2.384738	17943	2.746578629
408	2	3.537463	-0.182040265	2.689716	17937	1.406383112
408	3	3.32537	0.051645075	2.408833	17917	2.255547962
607	1	1.979009	0.405436722	1.4704	21281	2.376508455
607	2	1.624897	0.213823203	1.596135	21266	1.800484707
607	3	1.622589	0.603387547	1.013216	21252	1.663872909
708	1	2.742063	-0.294769277	2.329363	18123	2.11022555
708	2	2.605212	0.03281048	2.784566	18098	1.367081172
708	3	3.373201	-0.154129123	2.123944	18072	4.20949064
505	1	2.069387	0.827549303	2.001841	2008	1.656266385
505	2	1.725896	0.596510569	1.728748	2008	1.53669588
505	3	1.717482	0.620740606	2.031922	2008	1.38651187
507	1	2.011969	0.731406689	1.606235	1676	1.188399509
507	2	1.824898	0.718274379	1.599754	1676	1.092670314
507	3	1.522861	1.065859386	1.697011	1676	1.254368581

Notes. Parameter values of the sigmoid function used in the transformation from skill level estimates to points. *f* denotes the median and *e* and *k* is the slope of the curve just below and above the median, respectively. We are able to replicate all but 515 (0.03%) transformations from the raw test results to points. Source: UNI-C.

Appendix B. Descriptive statistics of raw test results

This appendix reports descriptive statistics of the distribution of test results by test year, grade and profile area. It is based on the raw test results (theta) supplied in data. Additionally, a test for mean of test score distribution equal to the mean of the normal distribution (H_0: mean equal to zero) is performed by a student's t-test. Mean equal to zero is rejected across all test scores distributions (omitted here).

Table B.1 Descriptive statistics of raw test results (theta) in reading, by test year, fagid, and profile area

Test year	Grade	Fagid	Profile area	Mean	Std. Deviation	Median
2010	2	102	P1	1.06	1.06	1.02
2010	4	104	P1	0.12	1.00	0.22
2010	6	106	P1	-0.05	0.92	0.02
2010	8	108	P1	0.70	0.90	0.78
2010	2	102	P2	1.72	1.71	1.72
2010	4	104	P2	2.15	1.25	2.15
2010	6	106	P2	2.58	1.26	2.65
2010	8	108	P2	3.99	1.46	4.22
2010	2	102	P3	0.48	1.59	0.56
2010	4	104	P3	0.55	1.16	0.63
2010	6	106	P3	0.79	1.00	0.80
2010	8	108	P3	0.90	1.08	0.96
2011	2	102	P1	1.17	1.07	1.13
2011	4	104	P1	0.21	0.99	0.30
2011	6	106	P1	0.03	0.94	0.10
2011	8	108	P1	0.76	0.93	0.80
2011	2	102	P2	2.01	1.70	2.12
2011	4	104	P2	2.25	1.24	2.25
2011	6	106	P2	2.72	1.28	2.79
2011	8	108	P2	4.15	1.46	4.31
2011	2	102	P3	0.62	1.52	0.75
2011	4	104	P3	0.81	1.12	0.83
2011	6	106	P3	0.96	1.05	0.95
2011	8	108	P3	1.08	1.08	1.13
2012	2	102	P1	1.23	1.12	1.18
2012	4	104	P1	0.14	0.99	0.25
2012	6	106	P1	0.08	0.94	0.15
2012	8	108	P1	0.83	0.93	0.84
2012	2	102	P2	2.07	1.70	2.21
2012	4	104	P2	2.06	1.28	2.24
2012	6	106	P2	2.80	1.32	2.86
2012	8	108	P2	4.53	1.50	4.62
2012	2	102	P3	0.68	1.50	0.82
2012	4	104	P3	0.85	1.14	0.88

2012	6	106	P3	1.01	1.07	0.98
2012	8	108	P3	1.18	1.08	1.24
2013	2	102	P1	1.29	1.08	1.25
2013	4	104	P1	0.23	0.99	0.33
2013	6	106	P1	0.15	0.97	0.19
2013	8	108	P1	0.92	0.98	0.92
2013	2	102	P2	2.16	1.69	2.32
2013	4	104	P2	2.20	1.28	2.33
2013	6	106	P2	2.86	1.34	2.92
2013	8	108	P2	4.62	1.49	4.71
2013	2	102	P3	0.75	1.46	0.91
2013	4	104	P3	0.95	1.14	0.96
2013	6	106	P3	1.07	1.08	1.04
2013	8	108	P3	1.23	1.07	1.29

Notes. The sample includes all mandatory and sick test observations of reading in the 2nd, 4th and 6th grade. Test results of pupils in special needs education (specialskoler og dagbehandlingstilbud/hjem) are excluded. Fagid refers to the variable supplied in data.

Table B.2 Descriptive statistics of the raw test results (theta) in mathematics, by test year, fagid, and profile area (p1, p2, p3)

Test year	Grade	Fagid	Profile area	Mean	Std_Deviation	Median
2010	3	203	P1	0.68	0.87	0.70
2010	6	206	P1	0.49	0.85	0.42
2010	3	203	P2	0.61	0.84	0.56
2010	6	206	P2	0.10	0.71	0.07
2010	3	203	P3	0.84	1.10	0.85
2010	6	206	P3	-0.15	0.85	-0.22
2011	3	203	P1	0.63	0.89	0.62
2011	6	206	P1	0.58	0.87	0.46
2011	3	203	P2	0.64	0.87	0.62
2011	6	206	P2	0.20	0.76	0.20
2011	3	203	P3	0.90	1.09	0.89
2011	6	206	P3	0.07	0.96	0.01
2012	3	203	P1	0.69	0.90	0.68
2012	6	206	P1	0.56	0.90	0.47
2012	3	203	P2	0.69	0.88	0.66
2012	6	206	P2	0.22	0.78	0.22
2012	3	203	P3	0.78	1.09	0.79
2012	6	206	P3	0.10	0.98	0.03
2013	3	203	P1	0.72	0.93	0.74
2013	6	206	P1	0.61	0.94	0.50
2013	3	203	P2	0.72	0.89	0.69
2013	6	206	P2	0.21	0.84	0.17

2013	3	203	P3	0.83	1.10	0.85
2013	6	206	P3	0.18	1.00	0.10

Notes. The sample includes all mandatory and sick test observations of mathematics in the 3rd and 6th grade. Test results of pupils in special needs education (specialskoler og dagbehandlingstilbud/hjem) are excluded.

Tables B.1 and B.2 support our concerns that the raw test score distribution within each profile area varies considerably and are not directly comparable. E.g. within the 2nd grade in 2010 the mean *(std. dev.)* of each of the three profile area is 1.05 *(1.07)*, 1.70 *(1.72)* and 0.47 *(1.60)*, respectively. Therefore, taking some simple average across profile area within each subject is not recommended.

Appendix C. Standardizing theta

The standardization is conducted accordingly:

i) Standardize theta with mean zero and standard deviation one and call it "zscore". This is done for each profile area, test and year resulting in three new variables (zscore_p1, zscore_p2, zscore_p3).

$$zscore_{pi_n} = \frac{theta_{pi_n} - \mu_{theta_{pi}}}{\sigma_{theta_{pi}}}$$

Where $i = 1,2,3$ denotes the profile areas of test observation n. Information of μ and σ within profile areas and years for reading can be found in Tables B.1 and B.2. This unit can be used directly for analysis, or to construct a measure of average skills within a subject.

ii) Then the average of the three profile areas is calculated for each pupil within each subject and year. Denote this by zscore_p123:

$$zscore_{p123\_n} = (zscore_{p1\_n} + zscore_{p2\_n} + zscore_{p3\_n})/3$$

This measure will have mean zero but the variance will be less than one.

iii) Therefore, the zscore_p123 is standardized once again within test and year and with mean zero and standard deviation one, resulting in the final measure of proficiency within a given subject: ztheta_p123.

Appendix D. Sample means

Table D. 1 Sample means and sample size of the full sample

	Full sample	
	Mean	Std. Dev.
Standardized test result (N=1,959,247)	0.029	0.971
Missing test results	0.074	
Exempted (N =1,600,238)	0.002	
9th grade exit examination marks		
Danish, reading	6.385	3.055
Danish,, spelling	6.400	3.072
Danish, essay	6.330	3.060
Danish, oral	7.422	3.575
Math, calculus	6.202	3.690
Math, problemsolving	7.334	3.577
English, oral	6.145	3.290
Physics/chemistry, oral	6.926	3.288
Pupil characteristics		
Girl	0.495	
Western immigrant/descendant	0.010	
Non-Western immigrant/descendant	0.090	
Low birthweight (<2500 g)	0.106	
First-born	0.431	
Second-born	0.372	
Third-born or later	0.187	
Multiple borns (e.g. twins)	0.038	
Born in the first quarter	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	0.018	
ADHD diagnosed	0.003	
School transfer within the last 2 years	0.191	
Assigned to special needs education, cause		
- learning siability	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	0.148	

Table D. 2 (continued) Sample means and sample size of the full sample

Familiy information (year 5)		
Mother's logearnings	9.634	4.854
Mother has negative earnings	0.166	
Mother's age	32.639	7.544
Mother's education:		
None or missing	0.051	
≤ High school	0.271	
Vocational	0.361	
Bachelor	0.246	
Higher	0.071	
Father's logearnings	10.275	4.800
Father has negative earnings	0.129	
Father's age	34.693	9.406
Father's education:		
None or missing	0.066	
≤ High school	0.248	
Vocational	0.412	
Bachelor	0.181	
Higher	0.093	
School information		
School size	465.461	177.810
Class size	21.729	3.953
Capital area school	0.064	
Bigger city school	0.115	
Absence information (N =1,569,223)		
Share of sick days	0.033	0.049
Share of absence	0.049	0.053
Has had legal absence	0.355	
Has had illegal absence	0.057	
N	2,116,150	

Appendix E. Correlations across profile areas

In general, the correlations between the raw test results are relatively high across profile areas in the same test (> 0.55). Table E.1 presents the raw correlation matrices of the 2010 reading and math tests. Although high, the correlation coefficients are seldom above 0.75, which can be interpreted as evidence of tests measuring a very specific set of skills within each profile area. Taking pupil background characteristics into account reduce correlation coefficients with around 0.03.

Table E. 1 Correlations between each of the three profile areas of the 2010 reading and math tests

Profile areas	P1	P2	P3	P1	P2	P3	P1	P2	P3
	Reading, grade 2			Reading, grade 4			Math, grade 3		
P1	1.00	0.59	0.59	1.00	0.66	0.71	1.00	0.64	0.72
P2	-	1.00	0.81	-	1.00	0.70	-	1.00	0.66
P3	-	-	1.00	-	-	1.00	-	-	1.00
	Reading, grade 6			Reading, grade 8			Math, grade 6		
P1	1.00	0.62	0.65	1.00	0.55	0.64	1.00	0.55	0.64
P2	-	1.00	0.69	-	1.00	0.59	-	1.00	0.59
P3	-	-	1.00	-	-	1.00	-	-	1.00

Notes. The table shows the raw correlations between the test results of each profile area within the 2nd grade, 4th, 6th and 8th grade reading test of 2010 and the 3rd and 6th grade math test. P1 denotes the raw test score of profile area 1 etc. The sample includes all pupils in public schools who have taken a national test in 2010.

Because we have chosen to use a standardized average of the standardized test results within the three profile areas of each subject, it may be of importance how the ranking of this average measure relates to the rankings of the individual profile area measures.

In general, our standardized average test score is very highly correlated with the standardized test results within each profile area of the subject, with coefficients between 0.82 and 0.92. Also, the average measure seems generally to be slightly higher correlated with the standardized test result of profile area 3 (e.g. for the 4th and 6th grade reading test, the 6th grade math test etc.), however, the difference is often minuscule and insignificant.

Table E.2 Correlations between standardized test results of each profile areas and the average standardized test score.

	P1	P2	P3	P1	P2	P3
	Reading, grade 2			Reading, 4th grade		
Standardized measure	0.8204	0.9020	0.9066	0.8779	0.8760	0.8959
	Reading, grade 6			Reading, 8th grade		
Standardized measure	0.8541	0.8738	0.8839	0.8406	0.8338	0.8312

Notes. The table shows the raw correlations between the test results of each profile area (P1, P2, P3) within the 2nd, 4th, 6th and 8th reading tests and the standardized average test score across the three profile areas. P1 denotes the standardized test result of profile area 1 etc. Standardized measures are calculated as described in section 5.1. The sample includes all pupils in public schools who have taken a national test in reading.

Appendix F. (AU internal)

This appendix is supplementary for AU internal researchers with access to data. We have chosen to share our programs, data and notes in a common folder available at project number 702727. Please contact the authors for access. Copying programs and data is at your own responsibility.

In this appendix we first describe where to find the available raw data. Second, we describe what can be found in the common folder, among other descriptive statistics presented in this paper.

Available raw-data

Raw Test results are found on the G drive in the following folders (DNT is an abbreviation of 'De National Test')

Mandatory test data:

- DNT 2010, 2011: 'F:\Rawdata\702727\data201204\dnt_au.sas7bdat'
- DNT 2012: 'F:\Rawdata\702727\data201211\dnt_au_foraar2012.sas7bdat'
- DNT 2013: 'F:\Rawdata\702727\data201401\dnt_au_foraar2013.sas7bdat'

Please note that some variables are changed in dnt2012 (and onward) compared to the earlier data.

- Dnt2010 and dnt2011: The variable "Skoletype" corresponds to the valuelist of DST's "inst2", i.e. it is a 3 digit classification of the schooling type.
- Dnt2012: The variable "Skoletype" corresponds to the valuelist of DST's "inst3", i.e. it is a 4-digit classification of schooling type. This implies a higher level of detail.
- Dnt2013: The variable "fagid" is no longer included; use the text-variable "fag".

Voluntary test data (i.e. either private schools or 'practice' tests):

- DNT voluntary 2010, 2011: 'F:\Rawdata\702727\data201211\dnt_au_frivillig.sas7bdat'
- DNT voluntary 2012: 'F:\Rawdata\702727\data201302\dnt_au_frivillig12.sas7bdat'
- DNT voluntary 2013: 'F:\Rawdata\702727\data201402\dnt_au_efteraar2013_v2.sas7bdat'

Exemption from the mandatory tests in 2010-2012:

- F:\Rawdata\702727\data201212\Dnt_evaluering_fritaget

Information of pupils' 9th grade examination and yearly marks for the school years of 2001/2002-2011/2012:

- F:\Rawdata\702727\data201302\tot_udfk2012

Economics Working Papers

2014-12: Louise Voldby Beuchert, Maria Knoth Humlum and Rune Vejlin: The Length of Maternity Leave and Family Health

2014-13: Julia Bredtmann, Sebastian Otten and Christian Rulff: Husband's Unemployment and Wife's Labor Supply – The Added Worker Effect across Europe

2014-14: Andrew B. Bernard, Valerie Smeets and Frederic Warzynski: Rethinking Deindustrialization

2014-15: Bo Sandemann Rasmussen: An Interpretation of the Gini Coefficient in a Stiglitz Two-Type Optimal Tax Problem

2014-16: A. R. Lamorgese, A. Linarello and Frederic Warzynski: Free Trade Agreements and Firm-Product Markups in Chilean Manufacturing

2014-17: Kristine Vasiljeva: On the importance of macroeconomic factors for the foreign student's decision to stay in the host country

2014-18: Ritwik Banerjee: On the Interpretation of Bribery in a Laboratory Corruption Game: Moral Frames and Social Norms

2014-19: Ritwik Banerjee and Nabanita Datta Gupta: Awareness programs and change in taste-based caste prejudice

2014-20: Jos Jansen and Andreas Pollak: Strategic Disclosure of Demand Information by Duopolists: Theory and Experiment

2014-21: Wenjing Wang: Do specialists exit the firm outsourcing its R&D?

2014-22: Jannie H. G. Kristoffersen, Morten Visby Krægpøth, Helena Skyt Nielsen and Marianne Simonsen: Disruptive School Peers and Student Outcomes

2014-23: Erik Strøjer Madsen and Yanqing Wu: Globalization of Brewing and Economies of Scale

2014-24: Niels-Hugo Blunch and Nabanita Datta Gupta: Social Networks and Health Knowledge in India: Who You Know or Who You Are?

2014-25: Louise Voldby Beuchert and Anne Brink Nandrup: The Danish National Tests – A Practical Guide