THE DANISH
NATIONAL CENTRE
FOR SOCIAL RESEARCH

Beatrice Schindler Rangvid

07:2014 WORKINGPAPER

Systematic Differences across Evaluation Schemes and Educational Choice

SFI – THE DANISH NATIONAL CENTRE FOR SOCIAL RESEARCH

SYSTEMATIC DIFFERENCES ACROSS EVALUATION SCHEMES AND EDUCATIONAL CHOICE

Beatrice Schindler Rangvid

THE DANISH NATIONAL CENTRE FOR SOCIAL RESEARCH, COPENHAGEN, DENMARK

Working Paper 07:2014

Systematic Differences across Evaluation Schemes and Educational Choice[*]

This version: November 2014

Beatrice Schindler RANGVID, The Danish National Centre for Social Research (SFI)

Abstract

Using large scale register data from Denmark in a difference-in-differences framework, I analyze whether systematic disparities between internal teacher scores and external exam scores in the school leaving certificates are linked to pupil characteristics. Such differences may be particularly consequential in a school system like the Danish, where post-compulsory education choices are made on ability signals only from teacher scores, as external assessments are not available until *after* these choices are made. I document that educationally disadvantaged groups (boys, low-SES, and migrant pupils) receive systematically lower teacher scores for equal exam scores than their advantaged peers. Using sibling fixed effects methods, I then simulate changes in educational choices for disadvantaged groups were they graded by their teachers as their advantaged peers. The results show an increase in low-SES pupils' predicted probability to enroll in high-school, closing almost 10% of the high-school enrolment gap to high-SES pupils. Increases for boys and migrant pupils are modest.

Keywords: educational economics; efficiency; difference-in-differences; education; grading gaps; sibling fixed effects; subjective assessment

JEL Classifications: I20

[*] Contact information: bsr@sfi.dk, SFI, Herluf Trolles Gade 11, DK-1052 Copenhagen K, phone: +45 3348 0884. Part of the research was conducted while working at AKF, Danish Institute of Governmental Research. I appreciate funding from the Danish Council for Strategic Research. The usual disclaimer applies. Thanks to Petter Lundborg, Paul Bingley, Arnaud Chevalier and participants at the EALE 2013 conference and the Meeting of the Strategic Research Advisory Board, The Danish National Centre for Social research (June 2013), for valuable advice.

1. Introduction

Countries all over the world use academic assessments to monitor pupil performance. Many different assessment procedures are in use in different parts of the educational system and across countries and their reliability and validity have received a great deal of research attention (e.g. Baird 2010). Disparities by type of assessment are potentially important and their impact on educational outcomes depends on the education system they are part of. In institutional settings using different types of assessment, they merely introduce uncertainty concerning pupils' academic ability. Yet, in settings relying on only *one* type of assessment, they may produce systematical divergence in the evaluation of pupils' academic ability. Among those groups of pupils whose skills are evaluated less favourably, this may lead to less ambitious educational choices, and, consequently, lower educational attainment.

Recent empirical work has shown that subjective assessment of pupils' academic abilities (e.g. teacher scores) often differ from (more) objective assessment methods like exam or test results in ways that are systematically related to gender and ethnicity, the gender and ethnic gap often being smaller with objective assessment procedures (e.g. Lavy 2008, Burgess & Greaves 2013, Cornwell et al. 2013, Gibbons & Chevalier 2011, Falch & Naper 2011, Lindahl 2007). These findings imply that some groups of pupils may be educationally disadvantaged simply by the *type* of assessment.

While disparities across different types of assessments seem to be a common trait in a variety of school systems, their potential *impact* on educational outcomes will vary across school systems depending on the way different types of assessments are used and their timing in relation to education choices. Some school systems pass information from more than one type of assessment (e.g. both teacher and exam scores/test scores) to pupils, teachers and parents at different stages of schooling. For example, in the UK, pupils traditionally take both external exams at various stages of schooling and receive teacher scores, and have thus been repeatedly informed about their academic ability by different types of assessment at the time when making decisions about post-compulsory education choices.[1] In these settings, grading disparities between different types of evaluations will mainly introduce *uncertainty* about pupils' ability, but are less likely to lead to systematically different feed-back on academic abilities. Yet, in other school systems, external assessments are administered only as school leaving exams, such that only teacher assessments are available towards the end of compulsory schooling, when pupils make crucial educational choices.

[1] Traditionally, pupils have been assessed by National Curriculum test by the end of Key Stage 2 (age 11) and Key Stage 3 (age 14). Yet, in 2009 testing at age 14 was abolished.

In such settings, systematic differences between teacher and exam scores potentially play a larger role, as this is the only measure of academic performance available to pupils, teachers and parents at the time of application for post-compulsory education. Here, systematic differences between teacher- and exam scores across groups of pupils may lead to less ambitious choices regarding post-compulsory education for pupils disadvantaged by teacher grading - and these groups may end up being underrepresented in subsequent education. This study therefore contributes to the debate around pupils' educational performance disparities by gender, SES and migration background, and the potential implication of this for their future life chances.

In this paper, I analyze systematic grading disparities and their potential consequences for post-compulsory education choices in a setting as described above. More specifically, I consider grading gaps at the end of lower secondary education in Denmark. Pupils' academic performance is evaluated only by their teachers until the end of lower secondary school[2]. The first formal external assessment is the school leaving examinations at the end of grade 9.

Grading methods can vary across a range of dimensions. They may be blind or non-blind, where blind grading refers to the pupil's identity not being known to the exam grader. Or, they may be subjective or objective, where objective refers to grading by an external examiner, while subjective grading is done by a teacher who knows the student from class. Teacher grades are obviously both subjective and non-blind. Exam grades in Denmark are non-blind, and partly objective since they are set jointly by the subject teacher and an external examiner. Final teacher grades are set just before the beginning of the exam period.

While the application procedure for post-compulsory schooling is completed in March of grade 9, results from the school leaving exams are only available by the end of June. Assessment scores seldom rigorously bar entry to upper secondary tracks in Denmark. However, the assessment of pupils' academic ability nonetheless influences those involved in the decision process (pupils, parents, teachers, school counselors) whether to enrol in high school, in vocational education and training (VET) or to drop out of the education system.

This study contributes to the literature being the first to examine the potential consequences of grading gaps in an education system, where only results from subjective assessments are available to decision makers at the time of making crucial educational choices. Thus, this is the first study to

[2] In 2010, a national test system has been introduced, providing standardized testing throughout the earlier years. The actual role these tests play in informing pupils, parents and teachers has not yet been examined.

provide an estimation of the potential ‘costs’ of this type of evaluation scheme for educational attainment. Moreover, most existing studies focus either on gender or ethnicity. In a unified framework, this study examines both dimensions and adds socio-economic status (SES) as a third dimension.

Contrary to most of the existing literature on grading disparities, this study does not consider the *sources* of such differences. This study endeavours to document grading disparities and analyse their potential consequences for educational choice of disadvantaged groups. If the results suggest that such detrimental consequences exist, a natural next step is to shed light on the underlying processes to suggest effective remedies[3].

The analysis is based on large-scale observational data from Danish administrative registers. In the main analysis, I use the population for the seven cohorts of pupils who completed grade 9 in 2005-2011, about 400,000 young people. First, I examine whether differences between teacher and exam scores vary systematically by pupil characteristics (gender, SES and migration background). I use a difference-in-differences framework (across type of assessment on the one hand and across gender, SES and ethnicity on the other hand) to estimate these grading disparities, exploiting the data structure with two scores for each subject, a teacher score and an exam score. After documenting the existence of grading disparities, I simulate how choice probabilities for post-compulsory education would change for pupils disadvantaged by teacher grading, if they were rewarded equally by their teachers as their advantaged peers (with equal exam performance).

I find that boys, migrants and pupils from low socio-economic background are systematically lower assessed by teacher scores than girls, natives and high-SES pupils with equal exam performance. Disparities in grading by gender and migration status are of similar size (0.13-0.16 SD), and for migrants are largely explained by differences in family background compared to natives. Grading differences by SES are much larger - 0.32 SD – yet, about half of the gap can be referred to differences in (other) family background characteristics. Using sibling fixed effects methods to account for differences in time-constant parental attitudes and preferences that are thought to influence both achievement and educational choice, my simulation results suggest that if low-SES pupils were rewarded equal teacher scores as their high-SES peers with equal exam performance, their predicted probability to enroll in high-school increases by 11%, closing almost 10% of the high-school enrolment gap to high-SES pupils. Increases for boys and migrant pupils are modest.

[3] This will be examined in another study.

In the remaining of the paper, I outline the background and related literature in the next section, and continue with a brief description of the Danish evaluation system. Section 4 presents the data, while sections 5, 6 and 7 show the results. The last section concludes.

2. Background and related literature

This paper adds to a small literature comparing different assessment methods in schools by pupil characteristics. Most studies explore differences by gender (Lavy, 2008; Cornwell et al., 2013; Falch & Naper, 2011), while Burgess & Greaves (2013) focus on ethnic minorities and Zavodny (2013) on overweight pupils. Most results document the existence of systematic grading gaps, favoring (educationally) advantaged pupil groups in more subjective assessments. Two other studies (Gibbons & Chevalier 2011, Kiss 2013) assess grading disparities along multiple dimensions (gender, migration, prior achievement).

While the existing literature centers on finding the *sources* of grading differences, two of the previous studies also investigate potential *consequences* of grading gaps for education outcomes. Burgess & Greaves (2013) show that the school-level test score gap between ethnic minority and white pupils is lower for the former in schools with large negative assessment gaps. Gibbons & Chevalier (2011) conclude that the divergence between the assessments at age 14 has almost no bearing on pupil qualifications or participation in education after age 16, and is unlikely to influence participation rates in higher education. Yet, unlike the present study, in the previous studies, pupils have received both teacher and exam scores *before* the education outcomes examined. Thus, Gibbons & Chevalier estimate the impact of *uncertainty* about own ability on educational outcomes, which may result from receiving differing scores in the teacher and exam assessments. In the present study, with exam scores unknown until *after* educational choices are made, I expect systematic grading differences to be more consequential.

Systematic divergence in assessment can arise through various channels that may affect either teacher scores or exam scores: teacher discrimination/stereotyping may have (more) influence in face-to-face assessments rather than in exam situations which are blindly - or at least externally - scored; teacher based and test based assessments may measure different skills (effort, non-cognitive skills), and competition in exam situations may be sources of grading differences. Yet, as argued above, while the sources of grading disparities are important to understand for policy advice on how to reduce them or at least mitigate their importance, this is not crucial in the context of the present study.

Less favorable teacher assessments (compared to exam scores) - in particular if this is the only ability measure available - may affect educational choice through a variety of mechanisms[4]. First, pupils' own academic perception is worse if less favourably assessed by the teacher, which might lead to less ambitious educational aspirations and choices. Second, educational guidance on choice of post-compulsory education is based on pupils' prior academic results – as assessed by the teacher. Thus, all important agents in making or guiding educational choice rely on information on pupils' academic skills only from teachers.

Third, less favorable teacher scores might also have an impact on educational opportunities due to more formal requirements in the admission to upper secondary education in the case of Denmark. To gain (direct) access to upper general education, pupils need an indication from their school on their 'readiness' to study the specific educational track they apply for[5]. This indication is given at the time of applying for post-compulsory education track, i.e. in the beginning of the final term of grade 9 (and prior to the school leaving exams). Moreover, lower teacher scores might make it harder to gain access to the preferred high-school, in particular in urban areas where pupils compete for a slot in sought-after high-schools. If admission to the preferred high-school depends on teacher scores[6], underassessment might lead to less-desired pupil-school matches, which in turn may affect dropout rates, achievement and final attainment.

Choice of educational track can be viewed in the framework of a rational choice model of educational decision making (Breen & Goldthorpe 1997). The model represents children and their parents as choosing among the different educational alternatives on the basis of their costs and benefits and of the perceived probabilities of success, i.e. as acting in a subjectively rational way. Lower teacher scores enter the model by altering the perceived probabilities of success in two ways. First, a specific educational track may only be open to pupils who meet some criterion, such as a

[4] There may, obviously, be other factors than academic ability that matter for educational choice. For example, students may differ in their comparative advantages or preferences with respect to manual labour vs. academic skills. An effect on educational choice cannot be ruled out.

[5] If he/she does not get this recommendation, the pupil can take an exam at the institution of the general upper general education to gain access, but still, this may deter some students from even trying.

[6] Formally, this should not be the case, but there is basis for discretion by the high-schools if schools are oversubscribed. High-schools are not formally allowed to make admission decisions on the basis of applicants' scores, but the student allocation procedures for oversubscribed schools make such behavior possible.

given level of ability measure, so that, for example, a pupil may only gain admission if his or her revealed ability level exceeds some threshold. Second, pupils' own perception of their ability are likely to shape their subjective probability of being successful in a specific educational track.

In this model, a lower measure of pupil ability may prevent pupils with ability levels just around the threshold of being admitted to a certain educational track from entering this track (at least without going through further testing). It might also lower pupils' subjective probability of success in alternative educational tracks, which may discourage pupils from choosing a more demanding track even if there is no formal bar to pass.

This kind of behaviour is documented in recent research, providing empirical evidence that high school and college students update their educational expectations based on information on their academic ability. Stinebrickner and Stinebrickner (2012) find that college students revise their expectations based on their actual score performance: the probability of drop-out rises when they perform significantly lower than expected. Zafar (2011) finds similar evidence of updating expectations, showing that students revise their expectations of future GPAs based on their prior GPAs, and Jacob and Wilder (2010) conclude that updating of educational expectations is based, in part, on the acquisition of new information about academic ability.

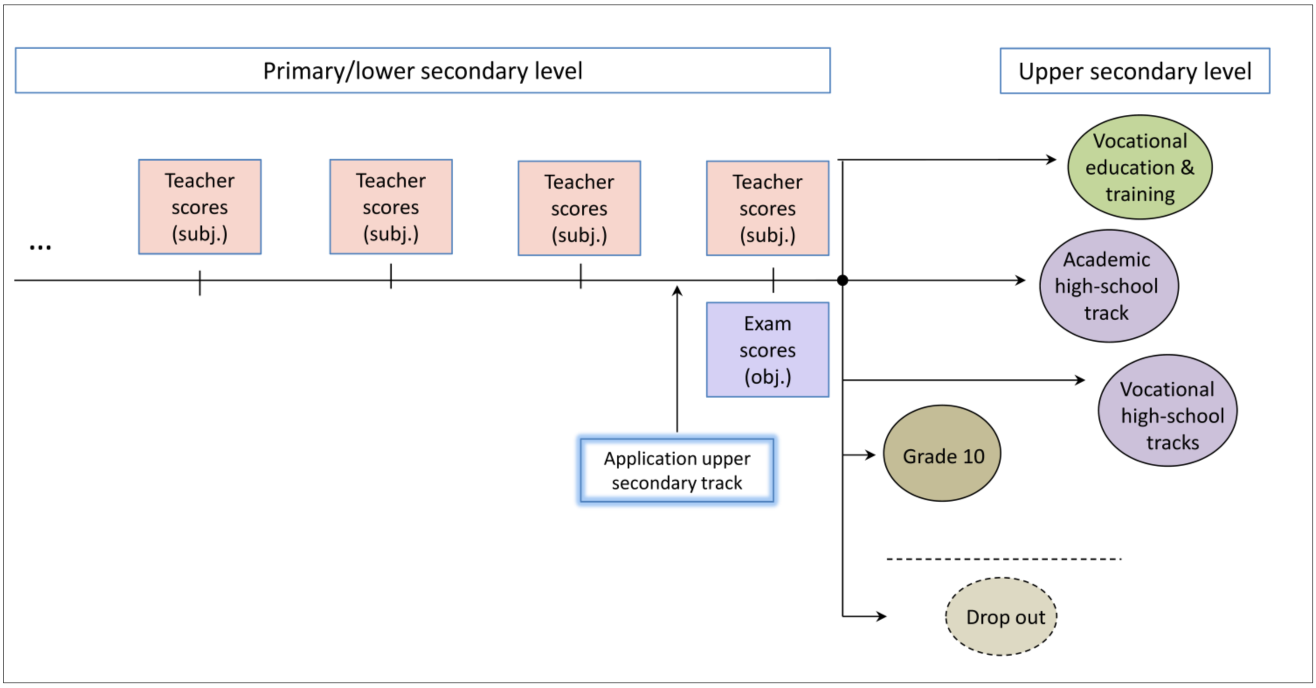
3. Danish institutional setting

Figure 1 illustrates the institutional setting in Denmark of the transition at the end of lower secondary education. Primary and lower secondary education consists of one pre-school year and 9 years of schooling in a comprehensive school system. No explicit streaming or tracking exist at the primary or lower secondary level. After grade 9, pupils continue in different education tracks, and a few drop out of the education system. A large share of pupils chooses to attend the intermediate, preparatory grade 10 before entering youth education programs. Grade 10 is optional and is chosen by a broad range of pupils, among them those undecided with respect to post-compulsory education choice and academically weak pupils advised to attend grade 10 to improve their results[7]. Another popular option after grade 9 is to enrol in different high-school tracks (academic or vocational), the academic high-school track being the most popular. Upon successful completion, all high-school

[7] After grade 10, the set of educational choice is the same as after grade 9.

tracks grant access to tertiary education. Other pupils enrol in vocational education and training (VET) straight after grade 9[8].

Figure 1: Institutional set-up.


Primary/lower secondary level
Upper secondary level
...
Teacher scores (subj.)
Teacher scores (subj.)
Teacher scores (subj.)
Teacher scores (subj.)
Exam scores (obj.)
Application upper secondary track
Vocational education & training
Academic high-school track
Vocational high-school tracks
Grade 10
Drop out

Throughout primary and lower secondary education, pupils and their parents repeatedly receive feed-back on academic performance from teachers. In public schools, teachers give verbal feed-back until grade 7, and only in grades 8 and 9 pupils receive teacher scores[9]. At the end of compulsory schooling in Denmark, pupils are evaluated by two sets of assessments that are included in the school leaving certificate: (1) scores awarded by the teacher in each subject, and (2) a set of mandatory final exams. Exam results are partly remotely scored as they are determined by the teacher and an external examiner, where the opinion of the external examiner dominates the teacher's opinion[10]. The purpose of using external examiners is to ensure that all pupils receive an unbiased assessment at the end of compulsory schooling that is comparable across schools (i.e.

[8] Yet, this is not the most common path to VET, as the large majority of pupils in VET complete the optional grade 10 before enrolling in VET.

[9] In private schools, different feed-back procedures may be employed.

[10] Note that not only are exams only *partly* remotely scored, but scoring is also non-blind, since the pupil's and school's name are stated on the exam paper for written exams. Oral exams are non-blind 'by nature'. Both features probably tend to *underestimate* the size of the estimated grading gaps in this study.

nationally consistent). Furthermore, the assessment of academic achievement must be based on educational objectives that are specified by the Ministry of Education for each subject (absolute grading) - grading may not be aimed at a particular distribution of scores (relative grading).

According to the 'Act on Public School', both teacher and exam assessments have the same stated purpose: to document the degree of compliance of pupils' skills with the objectives and requirements centrally specified for each subject. Thus, final teacher scores are intended to measure the *same set of skills* as the final exams. They are to be determined immediately before the exam period begins, such that results from pupils' performance at the final exams do not influence teacher scores. Also, they must reflect pupils' academic competence and skills at that point of time, implying that teacher scores are not meant to reflect coursework or pupil effort or other dimensions of achievement that are not purely academic. Thus, as both teacher and exam assessments are meant to measure the same set of skills, any systematic score differences across assessment schemes are therefore *not intended* by the law.

4. Data and variables

The data used in the empirical analysis is a dataset put together from different administrative registers hosted by Statistics Denmark. I work with different cohorts in the two parts of the study. In the grading gap regressions, I use seven cohorts of pupils (2005-2011). The data set contains information on roughly 400,000 pupils in public and private schools[11]. Data on pupil background is linked to assessment data via a personal registration number. The estimation sample includes only pupil-by-subject observations for which both teacher and exam scores are available.

For the second part of the empirical analysis (educational choice regressions and simulations), I need information on educational choice after grade 9, which is not yet available for the cohorts of 2010 and 2011. Therefore, I restrict the sample and run the educational choice models on the five cohorts of 2005-2009. For these cohorts, I have information on enrolment in education in autumn after completing lower secondary school. In the data, I both observe whether pupils enroll in education (or drop out) and also the specific educational program they enroll in.

[11] Including *efterskoler* (independent boarding schools for upper secondary students), but excluding schools for students with special needs.

Concerning the outcomes, both teacher and exam scores are reported on the same grading scale[12]. The dataset includes by subject-area teacher and exam scores for all pupils. The range of subjects included in the final examinations was changed as of 2007. In the years 2005-07, the subjects assessed by both exams and teacher grades are Danish (writing, spelling, neatness, oral); Math (written, oral, neatness); English, German/French (all oral); Biology and Physics/Chemistry. From 2008, subjects assessed by both exams and teacher grades include Danish (reading, writing, spelling, neatness, oral), Mathematics (proficiency and problem solving), English (written and oral), Physics/Chemistry, Biology, Geography, German/French, and adds a range of subjects that were not assessed by exams earlier on: History, Christian Studies and Social Studies[13,14,15].

I choose to include the full range of subject-areas assessed by both teacher grades and exams in both periods[16], because test scores for *all* subjects appear on the school report cards, and thus contribute to the overall picture of the pupil's academic ability to the pupil, the parents and school counsellors, providing signals of the pupils' ability that may matter for educational choices. All scores enter with equal weight in the empirical analysis.

I study grading gaps along three dimensions: gender, socio-economic status and migrant background. Pupils' socio-economic background is measured using parental education as a proxy. In Denmark, parental education is by far the most influential family background determinant of childrens' educational outcomes. I use a classification into three categories: none or only one parent has upper secondary education (high-school or VET) as the highest degree (low SES; 27% of pupils), both parents have some college education (high SES; 20%), and the remaining pupils are in

[12] The grading scale was changed in 2008. For use in the descriptive tables, 2005-07 scores are converted to the new scale. The possible scores in the new grading scale are 12/A, 10/B, 7/C, 4/D, 02/E, 00/Fx, -03/F. For use in the regressions, I standardize teacher and exam scores jointly within year and subject area to wipe out systematic differences across scales.

[13] Not all exams are administered to all pupils, though. For each class, the Ministry of Education draws one test subject from the science group (either Biology or Geography) and one from the Humanities (English (written), Christian studies (oral), History (oral), Social studies (oral) and German or French (written or oral). Tests in each of the five subjects within Humanities are drawn evenly across classes, such that each subject is covered by about 20% of students. A similar procedure applies to the two subjects within the Science group, which are taken by about 50% of the students each. See Table A1 for an overview.

[14] In some subjects, the teacher and exam score pair does not cover exactly the same form of assessment: in Physics/Chemistry, Biology, Geography, German/French, History, Christian Studies, Social Studies, teacher scores cover oral and written assessment in one single score, while exam assessments are either by written or oral exams (see Table A1). Yet, as I do not claim to be able to identify discrimination, whether the teacher and exam scores are strictly comparable is less important than the fact that all scores provide signals of the pupils' ability that may matter for educational choices.

[15] For the 2011 cohort, I exclude a small number of subjects where students have been assessed by the National tests in grade 8 (Danish, Biology, Geography, Physics/Chemistry). For these student cohort, results from an external assessment are available to the teacher before the final teacher grades are given. These external signals of the pupils' ability may influence teacher grades and these observations are therefore excluded from the analyses.

[16] I exclude a small number of subjects where exams are optional and chosen only by a small number of pupils.

the category with medium SES (53%)[17]. In the regressions, high SES is the reference category, i.e. grading gaps for low SES pupils compared to high SES pupils. I define immigrants as pupils whose both parents are born in non-Western countries[18], while the children themselves may be born in Denmark or in their country of origin. Third generation migrants and migrants from Western countries are included in the native category[19]. According to this definition, 9% of the pupils in the sample have a migrant background.

In the grading gap regressions, I control for family structure, the number of children in the family, parental income and occupation, and year of graduation (see Table A2). In the educational choice regressions, almost all controls are wiped out by the siblings fixed effects specification, the only remaining control being year of graduation.

5. Empirical strategy

The empirical analysis falls in two parts: first, I estimate grading disparities by gender, SES and migrant status between teacher and exam grades at the end of grade 9. Then, I use the estimated gaps in a simulation analysis to predict changes in disadvantaged pupils' post-compulsory education choices assuming that they are rewarded by their teachers as their comparator group.

In the first part of the empirical analysis, the structure of the data with two scores in each subject, one teacher-given and one exam score, allows the use of a difference-in-differences estimation strategy. Thus, I consider differences across type of assessment on the one hand and differences across gender, SES and ethnicity on the other hand. The specification is similar to Lavy's (2008) study for estimating gender gaps in Israel. To estimate systematic grading disparities between different groups of pupils, I regress the absolute difference between teacher- and exam-based assessments on indicators for gender, SES, and migration background, and the pupil level controls. Assuming linearity, the gap equation can be written as:

$$GR_{ijTA} - GR_{ijEX} = \alpha + \gamma \cdot \mathrm{Male}_i + \partial \cdot \mathrm{SES}_i + \delta \cdot \mathrm{Migrant}_{\ i} + \mu \cdot GR_{ijEX} + \boldsymbol{\beta}\mathbf{X}_\mathbf{i} + u_{ij} \qquad (1)$$

[17] Using an alternative definition with more narrow categories for low and high SES yields nearly identical results in the empirical analysis.

[18] Western countries are defined as EU-15, North America, Japan, Australia and New Zealand.

[19] Note that this might induce a downward bias in the estimated migrant grading gaps.

where GR_{iTA} og GR_{iEX} are the teacher score and exam score for pupil i in subject-area *j. Male*, *SES* and *Migrant* are the variables of interest: pupil gender, SES and migrant background. X_i is a vector of other pupil level controls, and u_{ij} is the residual. I also include the pupil's exam score to take account of the negative relationship between the teacher-/exam score gap and pupils' academic ability. Thereby, I seek to disentangle differences between groups that are solely attributable to their different locations on the achievement distribution from differences along social and ethnic lines[20]. This is equal to the specification in two recent articles by Burgess & Greaves (2013) and Cornwell et al. (2013). Furthermore, I add school fixed effects to account of differences in grading practices across schools. Remaining grading differences can thus only refer to different practices *within* the school[21].

In the second part of the empirical analysis - to gauge the importance of academic ability measures (as measured by teacher scores in this case) for post-compulsory education choices - I simulate educational choices of disadvantaged pupils, if they had been rewarded by their teachers like their advantaged comparator group with equal exam performance. I begin by estimating the importance of teacher scores for pupils' education choices[22]. As the choices are made in Spring *before* the end of grade 9, post-compulsory choice is based on the most recent teacher feed-back at that time, i.e. scores from the *first* term in grade 9. Yet, as this information is not available in the administrative registers, I approximate first term teacher scores by second term teacher scores (i.e. at the end of grade 9). Importantly, teachers determine the second term scores *before* the final exams, i.e. just as first term teacher scores, these are determined without knowledge of the student's performance at the final exam. It seems thus reasonable to assume that second (final) term teacher scores and first term scores are highly correlated – or at least not systematically different.

Educational choice is probably not only influenced by observable pupil and parents characteristics, but also by unobserved factors like parental ambition for their childrens education. To identify the

[20] When average teacher grades are plotted separately by exam scores, it becomes evident that teacher grading is compensatory, i.e. pupils with low exam performance have on average higher teacher scores (than their exam performance) and vice versa for pupils with high exam scores. Note, that teachers do not do this on purpose to compensate eg. for 'an exam that went bad', since teachers set teacher scores before students sit the exams. Alternatively, the observed pattern may be due to regression to the mean.

[21] As the data does not contain class or teacher information, I cannot do class or teacher fixed effects. In grade 9, pupils are typically taught by specialist subject teachers, at least in the core subjects. The number of classes per grade ranges from one to four, with most schools having two or three classes.

[22] Note, that I do not regress education choice on the difference in score between teachers and national exam in this analysis, since pupils do not know their level of performance at external assessments yet when making education choices.

model, I control for such factors that are common between siblings and constant over time by using sibling fixed effects models. The sibling sample is about one third of the total sample; yet, pupil and parent background characteristics in the sibling sample and the full sample are similar[23].

I use linear probability models[24] of two binary educational choices: (1) attending high-school vs. other choices and (2) attending the academic track of high-school education vs. other choices. In the regressions, I control for variables that vary at the sibling level (gender, year of graduation) and sibling fixed effects (θ_s). Controls that do not vary between siblings, as SES, migrant status, parental education etc. are wiped out by the sibling fixed effects specification. Formally, I write:

$$Educ_{is} = \alpha + \mu \cdot \mathrm{TA}_{is} + \gamma \cdot \mathrm{Male}_{is} + \delta \cdot \mathrm{gradyear}_{is} + \theta_s + u_{is} \qquad (2)$$

I estimate the effect of teacher scores on educational choices separately for each disadvantaged subgroup (males, low-SES & migrant background). I then use the estimated coefficients together with the estimated grading gaps in a simulation to predict changes in disadvantaged pupils' post-compulsory education choices assuming that they are rewarded by their teachers as their comparator group.

Caveats

Measuring 'true' disadvantage by teacher scores hinges on the availability of a valid measure of pupils' academic ability. In this study, this is approximated by exam scores, which gives us a (more) objective measure than teacher scores. However, two features of the exam scheme might make this approximation less than ideal. First, several exams are oral exams with both the teacher and an external examiner present. Oral (face-to-face) exams are probably more prone to systematic assessment differences than written exams. Second, one might be concerned that the non-blind grading procedure of written exams means that even scores from *written* exams are susceptible to systematic disparities (albeit to a lesser extent than teacher scores). Systematic grading differences are most likely to occur for the gender and migrant dimensions, which are easily revealed by the pupil's names. Two recent studies examine the effect of non-blind grading. Hinnerich (2011 a,b) conducts an experiment with blind grading of high-school exams in Sweden. The results show that while there is no gender difference between non-blind and blind grading of exams, there is a substantial disadvantage of 20 % of a standard deviation of the blind test score for ethnic minority

[23] Results are available on demand.

[24] As a robustness check, logit models have been run with similar results.

students from non-blind grading. This blind vs. non-blind difference is part of the 'true' grading disparity, but that cannot be measured in the present study, because blind exam grades are not available. Thus, the results in Hinnerich suggest that the migrant disadvantage in teacher grading in Denmark might be even larger than estimated in the present study.

While I cannot account for these flaws, it is essential to realize that they tend to underestimate, rather than overestimate, grading disparities. Thus, the results in this paper can be viewed as lower bound estimates of the true grading differences.

6. Grading disparities: empirical results

Table 1 presents descriptives on the score data. The table presents results for subgroups along the three dimensions of interest: gender, SES and migration status. To compare teacher scores of pupils with similar exam performance, we consider average teacher scores for pupils with an exam score of 7 (slightly above-average). The results show that male, low-SES and migrant pupils have lower teacher scores than their comparator group. For example, boys with an exam score of 7 are rewarded with average teacher grades of 6.61 compared to 7.09 for girls – a difference larger than one fifth of a standard deviation.

Table 1: Descriptive statistics by subsamples.

Exam score=7		Boys	Girls	LowSES	HighSES	Migrants	Natives
Teacher scores	*Mean*	6.61	7.09	6.40	7.37	6.51	6.89
	SD	2.13	2.02	2.14	1.99	2.21	2.08
Sample size	No. scores	762,558	838,901	387,618	335,080	103,127	1,498,332

I now turn to model these differences in a multivariate setting to isolate the respective contributions of gender, SES and migrant status. I run multivariate regression models (equation (1) of section 5) to examine the separate contributions of achievement levels and pupil characteristics. Scores for each subject-area have been standardized to a distribution with a mean of zero and a standard deviation of one. The standardization was applied to teacher and exam scores seperately within each year and subject-area. Standard errors are adjusted for clustering at the school level.

Table 2. Estimated grading gaps by gender, socio-economic status and migrant background

		Boys	LowSES	Migrants
(1) Raw gap	*Coef*	-0.090***	0.026***	0.077***
	se	(0.002)	(0.003)	(0.005)
	Adjusted R^2	*0,003*	*0.000*	*0,001*
(2) = (1) + ability	*Coef*	-0.156***	-0.328***	-0.133***
	se	(0.002)	(0.003)	(0.006)
	Adjusted R^2	*0.261*	*0.268*	*0.254*
(3) = (2) + other pupil background characteristics	*Coef*	-0.172***	-0.198***	-0.031***
	se	(0.002)	(0.003)	(0.005)
	Adjusted R^2		*0.290*	
(4) = (3) + SFE	*Coef*	-0.173***	-0.190***	-0.059***
	se	(0.002)	(0.003)	(0.004)
	Adjusted R^2		*0,300*	
	No. of (pupil/subject-area) observations		4.233.824	

Notes: * p<.05, ** p<.01, *** p<.001. Dependent variables are (gaps of) standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The number of observations is the number of exam takers times the number of subject areas, since the datasets are stacked (for each student there is one observation per subject area). Full results for the main specification (4) are available in Table A2.

I run four specifications of the model. In specification 1 and 2, I run three separate regressions to estimate gaps by gender, SES and migration background. Specification 1 does not include any controls, and in specification 2, I only add exam scores. In specifications 3 and 4, the gap coefficients are estimated in a joint model to take account of the correlation between eg. SES and migrant status, and I also add other pupil background controls. Specification 4 adds school fixed effects to specification 3. Table 2 shows results for the main parameters of interest: the estimated coefficients (and their standard errors) of the teacher-exam score gap, γ, ∂, and δ in equation (1)[25]. For example, a negative coefficient of -0.090 for the raw gender gap in Table 2 corresponds to a difference in the teacher/exam grading gap for boys of -0.09 SD of the score distribution relative to girls. In short, I term this a ‘grading gap against boys’ of 0.09 SD.

In specification 1 (‘raw gap’), I find a negative and statistically significant grading gap against boys. The gender gap is roughly one tenth of a standard deviation. The raw gaps for pupils from

[25] Full results for the main specification (specification 4) are available in Table A2.

low-SES backgrounds and pupils with migrant background are, against what one might expect, in favor of low-SES and migrant pupils. Yet, as shown in specification 2 (conditioning on exam scores), the positive sign of the raw gap is due to the low-SES and migrants' average location at the lower end of the achievement distribution, where teacher grading on average is more lenient compared to exam scores. As expected, also for boys the coefficient estimates decline due to the lower achievement relative to their comparator groups. With covariate adjustment in specification 3, the grading gaps by SES and migrant background decline (in absolute size) due to the correlation of SES and other covariates with migrant background. The gender gap changes only little, as background characteristics are similar for boys and girls. Specification 4 adds school fixed effects, i.e. the estimates now refer to within-school gaps. The gender and SES gaps are virtually unchanged by the inclusion of school fixed effects, suggesting that there are no systematic differences in grading gaps (e.g. due to different grading practices) across schools attended by boys and girls and by pupils with different socioeconomic backgrounds. Concerning the grading gap for migrant pupils, the overall estimate indicates that gaps may be slightly larger within schools than overall.

Overall, corrected grading gaps in specification 4 are largest by socioeconomic background (-0.190) and gender (-0.173)[26]. While the *additional* detrimental impact stemming from the migrant dimension is not large (-0.059 SD), migrant status tends to be correlated with low SES background, rendering the total grading gap for immigrant pupils larger than indicated by the estimate in specification 4. The results suggest that week performing groups - as measured by exam scores - like boys, low-SES and migrant pupils are evaluated even worse by teacher scores.

The results on grading disparities for gender can be compared to results from the studies by Lavy (2008) and Cornwell et al. (2013). If we average the by-subject results in Lavy (2008), the gender gap is -0.10 SD against boys, which is somewhat smaller than what I find. However, this would be expected as Lavy compares much more similar scores (two sets of exams) than this study. On the other hand, Cornwell et al. (2013) finds a gender gap against boys that is somewhat larger: -0.22 SD. All in all, the results document that grading disparities also exist in Denmark. They are credibly similar to those found in other countries, which is a good starting point to continue to the novel dimension of this study in the next section: assessing their potential consequences on pupils' subsequent choice of education.

[26] Note that the size of the SES gap will depend on the chosen definition of low vs. high SES background.

7. Grading bias and educational choice

After having documented the existence of grading disparities in the previous section, I now provide some evidence on the possible consequences of lower teacher scores for pupils' post-compulsory education choices. Low teacher scores might affect pupils' academic self-esteem, reducing his/her educational expectations if these choices are related to the feed-back on academic ability. Low teacher scores can be detrimental for academic attainment, if they change pupils' educational decisions toward less ambitious choices. I therefore hypothesize that if boys, migrants and low-SES pupils had received similar teacher scores as girls, natives and high-SES pupils with equal exam scores, they might have chosen more demanding post-compulsory education tracks.

To quantify the importance of academic ability measures (in this case, teacher scores) for post-compulsory education choices, I estimate the effect of teacher scores on the educational track that pupils enrol in after grade 9. I then continue by simulating how disadvantaged pupils' educational choices change, if they were rewarded by their teachers like their advantaged comparator group with equal exam performance.

Table 3. Education enrolment in year one after compulsory school and mean scores.

	High-school		*Academic high-school track*		Vocational education and training (VET)		Attending grade 10		Not enrolled		All pupils		
	% of sample	Mean TAscores	% of sample	Mean TAscores	% of sample	Mean TAscores	% of sample	Mean TAscores	% of sample	Mean TAscores	% of sample	# pupils	Mean TAscores
All	32%	7,9	*24%*	*8,1*	11%	4,6	48%	5,7	9%	5,2	100%	297.771	6,3
Boys	30%	7,8	*19%*	*8,0*	15%	4,4	47%	5,4	8%	4,7	50%	148.145	5,9
Low-SES	21%	7,5	*13%*	*7,6*	18%	4,3	50%	4,9	11%	4,2	28%	84.760	5,2
Migrant	35%	*7,2*	*25%*	*7,4*	13%	3,4	44%	*4,0*	8%	3,9	8%	24.920	5,1

Note: TA designates Teacher Assessment, EX is for Exam. 2005-09 cohorts.

Table 3 shows descriptive results on educational enrolment in the autumn following graduation from grade 9. Overall, 32% pupils enroll in high-school - 24% opt for the academic track. A slightly higher share of migrant pupils enrol in high-school and the academic track (35 and 25%), while these choices are less common among boys and low-SES pupils. Overall, 11% of pupils enroll in a vocational education and training (VET) program after grade 9. VET is a more popular choice among boys (15%) and low-SES pupils (18%). Almost one in two pupils enrolls in the optional grade 10. Grade 10 is slightly more popular among low-SES pupils (50%) and somewhat less common among immigrant pupils (44%). Overall, 9% of the young do not continue in education after compulsory school. GPA from the 9th grade school leaving certificate is highest among pupils attending high-school and lowest among those entering VET programs immediately after compulsory school. Thus, teacher assessed academic ability appears to be related to choice of educational track after grade 9.

To take account of potentially confounding factors, I estimate the effect of teacher scores on track choice using sibling fixed effects methods. Table 4 reports regression results of the estimation of equation (2). Coefficient estimates for teacher scores are displayed for the three educational choice outcomes and the three disadvantaged groups of pupils (boys, low SES pupils, migrants). The results suggest that teacher scores significantly affect educational choices in the expected direction. In particular for boys, a one standard deviation increase in teacher scores is related to an increase in attending high-school (compared to other choices) by 4.1 percentage points and in enrolling in the academic track by 2.6 percentage points.

Table 4. Impact of teacher scores on post-compulsory education choice (sibling fixed effects).

	Attending high-school		Attending academic high-school track		No. obs.
	Coef	se	Coef	se	
All	0.073***	0.000	0.053***	0.000	1,411,772
Boys	0.041***	0.000	0.026***	0.000	704,267
Low-SES	0.068***	0.001	0.045***	0.000	342,409
Migrant	0.104***	0.001	0.078***	0.001	132,749

Notes: Only within sibling-pair varying controls included. * p<.05, ** p<.01, *** p<.001. 2005-09 cohorts.

For low SES pupils, choice probabilities increase by 6.8 and 4.5 percentage points, respectively. Heterogeneous effects by pupil groups suggest that educational choices of migrant pupils are more affected by teacher scores than other pupil groups. Among migrants, a one standard deviation increase in teacher scores increases high-school enrolment and high-school academic track enrolment by 10.4 and 7.8 percentage points, respectively.

I proceed with the analysis by simulating how educational choices might change, if teachers had awarded boys, low-SES and migrant pupils similar teacher scores like their comparator groups with equal exam performance. The simulation is done in two steps. First, I predict teacher scores for the disadvantaged groups when they are remunerated by their teachers like their advantaged peers. In the second step, I use the predicted teacher scores together with the coefficients from the educational choice regressions to calculate the predicted changes in educational choice probabilities.

Table 5. Simulated changes in educational choice probabilities.

	High School			Academic HS track		
	Change in probability	Mean level	% change	Change in probability	Mean level	% change
Boys	0,007	0,30	2%	0,004	0,19	2%
Low-SES	0,023	0,21	11%	0,016	0,13	12%
Migrant	0,016	0,35	5%	0,012	0,25	5%

Note: Simulated changes in educational choice probabilities if disadvantaged groups had received equal teacher scores as their advantaged comparator group (conditional on exam performance) using coefficients from Table 2 & 4.

The simulation results are shown in Table 5 indicating that an increase in average teacher scores increase the probability of enrolling in more ambitious education tracks for all disadvantaged groups. The results suggest that the effect is most important for low-SES pupils and for the decision to enroll in high school straight after grade 9. If low-SES pupils were remunerated by teacher scores as their high-SES peers, this would increase their probability of enrolling in high-school by 2.3 percentage points. This corresponds to an increase of 11% off the actual enrolment share of 21%. Calculated as a percentage change to current enrolment, the predicted change of enrolling in the academic high-school track is equally important (a 12% increase off the 13% currently enrolling). These predicted changes in enrollment would close almost 10% of the general high-school enrollment gap to high-SES students, and 6% of the academic track enrolment gap[27]. The simulated changes for boys and migrant pupils are modest.

8. Sensitivity analyses

In this section, a few concerns regarding the robustness of the results are examined. To begin with, I provide some sensitivity checks concerning the functional form of the grading gap regression models of section 6. Then, I examine the robustness of the second stage simulations of section 7.

[27] The general high-school enrolment rate straight after grade 9 among high-SES pupils is 49%. 42% are enrolled in the academic track.

The first check concerns outliers, i.e. pupils performing very well or poorly in the exam situation, may be driving the results. When I limit the sample to pupils around the mean of the exam score distribution[28] and reestimate specification 4 from Table 2, I find that the male and low-SES disadvantage are somewhat larger in the restricted sample, while the migrant disadvantage is slightly smaller (Table 6, col. 2[29]). However, overall, the results are not substantially affected.

Second, since both the teacher and the exam scores come from a seven-point scale, the gap variable does not vary much, and a continuous specification might be unsuitable. I therefore reestimate the main specification using a linear probability model for the likelihood that the teacher score is equal to or greater than the exam score (similar to Burgess & Greaves 2013). The results (Table 6, col 3) show that boys are 6.1 percentage points less likely than girls to have teacher scores equal to or exceeding exam scores, while low-SES pupils are 6.2 percentage points less likely than high-SES pupils and migrant pupils are 2.5 percentage points less likely than natives. While the size of these coefficients cannot be directly compared to the main specification, I note that the signs and relative size of the gaps are unchanged.

In order to make sure that the results are not qualitatively influenced by schools with an overwhelmingly native pupil intake, I repeat the analysis excluding schools with fewer than five migrant pupils as suggested in Burgess & Greaves (2013). Imposing this restriction reduces the sample by almost 25%. However, the results are similar to the full sample results (Table 6, col 4).

[28] Here defined as achieving scores 4 and 7 on the non-standardized scale.

[29] The results from specification 4 of Table 2 are repeated in column 1 of Table 6 for comparison purposes.

Table 6. Robustness checks using different samples and specifications

	1	2	3	4
	Main specification	EX=4,7	TA>EX	Min. 5 migrants
Boys	-0.173***	-0.200***	-0.061***	-0.170***
	(0.002)	(0.002)	(0.001)	(0.002)
Low-SES	-0.190***	-0.209***	-0.062***	-0.192***
	(0.003)	(0.004)	(0.001)	(0.003)
Migrants	-0.059***	-0.045***	-0.025***	-0.055***
	(0.004)	(0.005)	(0.002)	(0.005)
Observations	4.233.824	1.760.673	4.233.824	3.194.381
Adjusted R^2	0,300	0,132	0,144	0,296

Notes: * p<.05, ** p<.01, *** p<.001. All controls included as in specification 4, Table 2. Dependent variables are (gaps of) standardized scores. Standard errors are corrected for school-level clustering and are presented in parentheses. The number of observations is the number of exam takers times the number of subject areas, since the datasets are stacked (for each student there is one observation per subject area). TA stands for Teacher Assessment, EX is for Exam. Specification (1) is the main specification as in Tb. 2, spec. (4). Specification (2) limits the sample to pupils with exam scores equal to 4 or 7. Specification (3) models the probability that TA>EX. The dependent variable is binary, equal to one if TA>KS. Specification (4) limits the sample to pupils with at least 5 migrants in the school by year cohort.

A last robustness check concerns the validity of the identification strategy of the educational choice regressions. In section 7, I use sibling fixed effects to account for differences in time-constant parental attitudes and preferences that are thought to influence both achievement and educational choice. Sibling fixed effects estimation is doing this quite efficiently - but only for the restricted sample of siblings in the data set. Pupils without any siblings, and pupils without siblings graduating from grade 9 among the 5 cohorts of our study are not included in the regression. As an additional identification strategy, I can - for the gender dimension[30] - use an alternative identification strategy to account for differences in parental attitudes and preferences. For a subsample, I have information from a survey on parents' attitudes and preferences regarding their child's education[31]. Parents are asked to what degree they consider it important that their child is

[30] Due to data restrictions, this check can by carried out only for the gender dimension.

[31] I use data from the Danish Longitudinal Childhood Study (DALCS; 2011-wave), which provides information from a survey to the parents. The survey initially includes 6,000 youngsters born in 1995. About 4000 answered the survey, and 75%, or 3,000, of these sat for their school leaving exams briefly thereafter (most of the remaining pupils attend grade 8 in 2011 and thus complete compulsory school the following

doing well at school and, generally, the importance they attach to education as a means to obtain a good economic and social position in the society. I include this information as controls in the educational choice regressions. As pupils in this subsample all are from the 2011-cohort, educational choice data from the administrative registers, is not available yet. However, as part of the data collection, pupils were asked about their choice of education for the following school year (i.e. the year after grade 9)[32]. We use this information to construct the same set of outcome variables as in section 7 and run the educational choice regressions. I cautiously conclude that the results do not provide any evidence that the importance of teacher scores is overestimated by the sibling fixed effects estimation in the male subsample; quite the contrary.

9. Conclusion

Differences in the timing and use of pupil assessment schemes across countries are likely to have an impact on how severe the potential consequences of grading disparities are for educational outcomes. This study extends the literature on pupil assessment schemes to examine grading gaps and the role they play in educational decisions in an institutional framework, where important choices are based on feed-back on academic ability only from teachers. The results show that certain groups of pupils are disadvantaged simply by the form of assessment. Boys, low-SES and migrant pupils with exam scores like their female, high-SES and native counterparts are systematically awarded lower teacher scores.

Simulating consequences of grading disparities for post-compulsory track choices, I find large changes in the probability to enrol in high-school for low-SES pupils: being rewarded with equal teacher scores as their high-SES peers increases high-school enrolment by 2.3 percentage points off an enrolment rate of 21%, closing roughly 10% of the high-school enrolment gap to high-SES peers. The estimated effects are probably lower bounds of the true effects, since grading disparities are probably underestimated due to the non-blind and only partly external nature of the exam scheme. Thus, I probably underestimate the true disadvantage of boys, migrant and lowSES pupils.

year). The survey sample is probably prone to selection bias as well, but in different dimensions, and thus provides a check of our main results.

[32] The survey was administered in Spring 2011 and the answers should thus be good proxies for pupils' final choices.

These results contribute to the debate of persisting gaps in education across demographic groups, and the potential implication of this for their future life chances. First, the results of this study question the concept of ‘equal opportunity’ in the Danish education system. Grading disparities against disadvantaged groups have been documented in various countries. However, the fact that teacher feed-back stands alone for most of the pupils school career is a feature in the Danish school system that leaves them a larger (adverse) influence than necessary.

Second, biased feed-back on academic ability can have other detrimental effects than the ones examined here. Systematically lower teacher scores could affect pupil learning by reducing pupil effort. Also, if lower teacher scores are related to lower teacher effort, this also may harm pupils’ learning and reduce the accumulation of skills.

Thus, these arguments and the results found in this study suggest that one-sided feed-back on academic ability from teachers should be avoided. Some recent changes in the Danish education system have the potential to work in this direction. First, schools can choose to administer computer based, automatically scored tests in some subjects (for the time being Geography and Biology) instead of paper and pencil tests. This makes a completely blind scoring procedure with no subjective judgements involved.

Second, beginning in 2010, national tests are administered to pupils in core subjects at various stages (from grade 2 to grade 8). Also these tests are computer based with automatically generated scores, thus meeting the criteria of an external and blind scoring procedure. Results are available to teachers, pupils and parents, and hold the potential to challenge teachers’ perception of pupils – perhaps bringing a decrease in grading disparities and achievement gaps. However, it is not yet clear how rigorously results from these tests are disseminated to pupils and parents, and the importance accorded to them. Unless these scores play an equally important role in the school system as teachers grades, their influence will be limited. The introduction of national tests in 2010 is an opportunity to assess the impact of earlier external signals of students’ ability. This is an avenue for future research.

Finally, the results caution against rolling back existing central test systems as recently done in the UK, where the Key Stage 3 national tests (when pupils are 14 years old) have been abolished in 2009. As Burgess & Greaves (2013) show, teacher grades disadvantage weak groups of pupils also in the UK, thus, the abolishment of the Key Stage 3 tests is hardly to their favour.

References

Baird, J.-A. (2010): ‘Examinations versus teacher assessments’, *Assessment in Education: Principles, Policy & Practice*, 17:3, 251-254

Breen, R. & J.H. Goldthorpe (1997): ‘Explaining educational differentials: Towards a formal rational action theory’. *Rationality and Society*; 9(3): 275-305.

Burgess, S.M. & E. Greaves (2013): ‘Test scores, subjective assessment and stereotyping of ethnic minorities’, *Journal of Labor Economics;* 31(3): 535 - 576.

Cornwell, C.; D. Mustard & J. Van Parys (2013): ‘Non-cognitive Skills and Gender Disparities in Test Scores and Teacher Assessments: Evidence from Primary School’, *Journal of Human Resources;* 48(1): 236-264.

Falch, T. & L. R. Naper (2011): *‘Educational Evaluation Schemes and Gender Gaps in Student Achievement’*. Working Paper No. 4/2011, Department of Economics, Norwegian University of Science and Technology (NTNU)

Gibbons, S. & A. Chevalier (2008): ’Assessment and age 16+ education participation’. *Research Papers in Education, 23*(2): 113-123.

Jacob, B.A. & T. Wilder (2010): ‘Educational Expectations and Attainment’. NBER Working Paper #15683.

Hinnerich, B.T., Höglin, E. & M. Johannesson (2011a): ‘Are boys discriminated in Swedish high schools?’ *Economics of Education Review*, 30 682–690.

Hinnerich, B.T., Höglin, E. & M. Johannesson (2011b): ‘*Ethnic Discrimination in High School Grading: Evidence from a Field Experiment*’.

Kiss, D. (2013): ‘Are immigrants and girls graded worse? Results of a matching approach’, *Education Economics*, 21(5): 447–463.

Lavy, V. (2008): ‘Do Gender Stereotypes Reduce Girls' Human Capital Outcomes? Evidence from a Natural Experiment’. *Journal of Public Economics*, 92: 2083-2105.

Lindahl, E. (2007): ‘*Comparing teachers' assessments and national test results - evidence from Sweden’*. Working Paper Series, No. 24. Institute for Evaluation of Labour Market and Education Policy.

Stinebrickner, T. & R. Stinebrickner (2012): 'Learning about Academic Ability and the College Dropout Decision', *Journal of Labor Economics,* University of Chicago Press, vol. 30(4), pages 707 - 748.

Zavodny, M. (2013): 'Does weight affect children's test scores and teacher assessments differently?' *Economics of Education Review,* 34: 135-145.

Table A1. Final assessments in grade 9

				Final examination		Percentage of students who sit the exam	Final teacher grade			Mandatory exams	Exact TA/EX match
				Written	Oral		Written	Oral	Written/oral		
(1) Mandatory core subject examinations		1	Danish, reading	X		100%	X			x	x
		2	Danish, writing	X		100%	X			x	x
		3	Danish, spelling	X		100%	X			x	x
		4	Danish, neatness	X		100%	X			x	x
		5	Danish, oral		X	100%		X		x	x
		6	Mathematics, skills/competences	X		100%	X			x	x
		7	Mathematics, problemsolving	X		100%	X			x	x
		8	English, oral		X	100%		X		x	x
		9	Physics/chemistry		X	100%			X	x	
(2) Mandatory randomly selected subject examinations	*Humanities*	10	English, written	X		20%	X			x	x
		11	2nd foreign language	(X)	X	20%	X	X		x	x
		12	History		X	20%			X	x	
		13	Social studies		X	20%			X	x	
		14	Christian studies		X	20%			X	x	
	Science	15	Biologi	X		50%			X	x	
		16	Geography	X		50%			X	x	

Note: The table describes the exam scheme as from 2007. For 2005 and 2006, we include grades for the following subjects: Danish (writing, spelling, neatness, oral); Math (written, oral, neatness); English, German/French (all oral); Biology, Physics/Chemistry. Furthermore, in 2007, exceptionally, the randomly selected subjects in humanities were all assessed by written exams. After that, only the English exam continued by written assessment, the others became oral exams.

Table A2. Full results for specification 4, Table 2

	Coef	se
Boys	-0.173***	(0.002)
Low-SES	-0.190***	(0.003)
Medium-SES	*-0.085****	*(0.002)*
High-SES	*Reference*	
Migrant	-0.059***	(0.004)
Exam score (std)	-0.467***	(0.001)
Broken family	-0.086***	(0.002)
1 child in family	-0.021***	(0.001)
2 children	*Reference*	
3 children	-0.002	(0.002)
4+ children	-0.027***	(0.003)
Income, mother	0.039	(0.027)
Income, father	0.038***	(0.006)
Mother: Self-employed	-0.012***	(0.003)
Mother: wage earner, top	0.047***	(0.003)
Mother: wage earner, medium	*Reference*	
Mother: Wage earner, bottom	-0.051***	(0.002)
Mother: Wage earner, other	-0.061***	(0.003)
Mother: Permanent income transfers	-0.107***	(0.005)
Mother: Others	-0.076***	(0.007)
Father: Self-employed	-0.034***	(0.003)
Father: wage earner, top	0.031***	(0.002)
Father: wage earner, medium	*Reference*	
Father: Wage earner, bottom	-0.064***	(0.002)
Father: Wage earner, other	-0.061***	(0.002)
Father: Permanent income transfers	-0.101***	(0.004)
Father: Others	-0.069***	(0.004)
Graduation cohort 2005	*Reference*	
Graduation cohort 2006	0.005	(0.004)
Graduation cohort 2007	0.006***	(0.004)
Graduation cohort 2008	-0.054***	(0.004)
Graduation cohort 2009	0.006	(0.004)
Graduation cohort 2010	-0.077***	(0.005)
Graduation cohort 2011	-0.013***	(0.005)
Constant	0.337***	(0.010)
Observations	4233824	
Adjusted R^2	0,300	

Note: Standard errors in parentheses * p<.05, ** p<.01, *** p<.001 Missing value flags for all variables are included in the regression, but not shown.