Teacher Discretion in Grading Standardized Exams: Stakes and Information in Dutch Secondary Education Exams*

Ilja Cornelisz
Chris Van Klaveren

Keywords: Teacher Discretion, Grading, Equity

1 Introduction

Standardized tests serve to provide objective measures for student performance and can be high stakes for students as they determine, at least in part, retention and graduation decisions (Dee et al., 2016). These standardized tests also have become increasingly central to accountability policies with the objective to evaluate, for example, school and teacher performance. The main intent of test-based accountability policies is to provide incentives that maximize student learning, but perverse incentives resulting from badly designed accountability policies can have significant, unintended and undesirable consequences (Jacob, 2005). The widespread concerns over test validity and the manipulation of scores are therefore not surprising (Dee et al., 2016), yet there is surprising little empirical evidence related to test-based accountability and how it may lead to manipulation of student test scores (Jacob, 2005). Two exceptions are two recent empirical studies that provide strong evidence that there is all the more reason for policy makers to be aware of the implications of teacher discretion in grading standardized exams.

Dee et al. (2016) examined the causes and consequences of test score manipulation of high-stakes exit exams for New York State secondary-school students and find that teachers

*Chris van Klaveren (c.p.b.j.van.klaveren@vu.nl) and Ilja Cornelisz (i.cornelisz@vu.nl) are both affiliated with Faculty of Behavioral and Movement Sciences at VU University Amsterdam and the Amsterdam Center for Learning Analytics (*ACLA*,acla.amsterdam).

purposefully moved students just over predefined performance thresholds when grading their own students. Moreover, results varied systematically across and within schools and had heterogeneous implications with respect to subsequent student outcomes. Notably, conditional on scoring near a proficiency cutoff, white and Asian students and students with better baseline scores and good behavioral records are more likely to benefit from such teacher discretion. Diamond and Persson (2016) corroborate the existence of test score manipulation for Swedish compulsory schools, and similarly identify "a bad test day"-effect, suggesting that teachers exploit their discretion to undo potentially harmful consequences of idiosyncratic student performance. Yet, in contrast, their estimates are not related to student background characteristics. Furthermore, they find relative homogenous positive implictations for subsequent educational, labor market and life outcomes, highlighting that signaling mechanisms might enhance a student's academic motivation and/or teachers' perception of academic ability.

This study adds to the emerging literature on local grading, teacher discretion and test score manipulation (see, also: Lavy, 2008; Hanna and Linden, 2012; Burgess and Greaves, 2013) by evaluating scores on high-stakes standardized exams at the end of secondary education in the Netherlands. A unique feature of the Dutch exam system is that subject teachers grade the standardized exam of some of their students twice over a short span of time, but in a vastly different context in terms of both the *stakes* at hand and the *information* available. Students are allowed to retake an exam for one subject and this re-exam takes place one week after the results of the first attempt have become known. Student generally retake an exam if the aggregate results across all subjects after the first term are insufficient for overall graduation (i.e. if they do not pass their exam). It follows that the stakes of the re-exam are even higher than the stakes of the first-attempt exames, since graduation depends on it. The information available to students and teachers in the first and second term is distinctively different, in the sense that both teachers and students know exactly how many points are needed to graduate in the second term, but not in the first term. This is caused by the fact that the Dutch Testing Agency (CITO) announces the conversion formula which must be used to transfer achieved points to grades after the first attempt *and,* generally, this conversion formula also applies for the second term. As a consequence we know for the re-exam (i.e. the second term) that (1) it is optimal for the students to perform as well as possible on the re-exam, and that (2) grade manipulation by teachers will particularly reveal in the second term.

We exploit this setting to evaluate the implications of stakes and information with respect

to the potential for teacher discretion effects in grading standardized exams. Using administrative data for the Netherlands for the period 2007-2014, we find that there is considerable bunching in the distribution of student test scores. The discontinuity is most notable just above the cutoff required to pass a particular subject. Disaggregating results by whether the score is obtained in the first or second term reveals that this discontinuity is completely driven by students who take up the opportunity to retake the exam for one subject. Results show that the student-body population of the retry exam overwhelmingly consists of those who failed to graduate based on results after the first term, thereby underscoring the high-stake nature of the retake exam. More importantly, the findings reveal considerable systematic variation in the size of this discontinuity, both across and within schools, and with estimates related to both student-, and subject characteristics.

In order to explain the mechanisms behind these findings, we relate variation in school-level discontinuities over time to whether or not a school was under the scrutiny of more intense supervision by the Education Inspectorate. Furthermore, we are able to relate results to the nature of test and find that exams with relatively many open-response or essay questions, as opposed to multiple-choice questions, display larger discontinuities around the cutoff. Finally, we explore what the potential consequences are for subsequent educational decisions and outcomes (e.g. higher educational enrollment). These results have important implications for educational policy in general, and for local grading of standardized exams in particular.

2 Context, Data and Methodology

The school-leaving examination for secondary education in the Netherlands consists, for each subject, of a school examination and a national written examination at the end of the final school year. Schools set their own school exams, but the Ministry of Education, Culture and Science prescribes which topics must be covered. School examination dates are not nationally fixed and the school exam usually comprises two or more tests per subject, which may be oral, practical or written. Subjects outside the national exam framework may be completed before the final school year. For subjects in the national exam framework, the (weighted) average score on the school examinations counts for 50 percent towards the overall result for a subject. Depending on the level of education, students take national exams in about 6-8 different subjects.

For these subjects, the remaining 50 percent is determined by the national exam. There

is one national written exam per subject for all students at the same level of education. The national exam for a subject always takes place at a fixed date and time at the end of the final year and is constructed by the Ministry of Education, Culture and Science. The grading scale is from 1 to 10, where 1 is the lowest and 10 the highest grade. A grade of 5.5 is required to pass a particular subject, but for school leaving examinations, where six or more subjects are examined, there are explicit rules by which insufficient grades for one or more subjects can be compensated by high grades in other subjects. If a student fails to graduate, (s)he has the opportunity to retake the exam for one subject in the so-called second term. This exam takes place approximately one week after the first-term results have become known. The highest score on both attempts is used towards determining whether a student has met the requirements for graduation. If that's not the case, a student will have to retake the final school year in its entirety the next year.

A relatively uncommon, but for this study crucial, feature of the Dutch secondary education system is that the national exams are graded by students' own subject teachers. Explicit guidelines for the grading procedure are provided. Based on these guidelines, teachers assign a score to each answer, after which the results per question are uploaded digitally. The exams are then send back to the Dutch National Institute for Educational Measurement (CITO), who is charged with the logistics surrounding the exam and who then assigns a teacher from a different school to re-mark the work (the so-called second corrector). After both colleagues agree on the number of points awarded to each student for each question, the scores are uploaded (digitally) to a central system. Importantly, it is only after these results have been handed in, that the CITO makes available the conversion formula necessary to be able to transfer the points given to a grade. This conversion formula contains a factor which varies from year to year as to control for erratic differences in the difficulty of a national exam. In practice, depending on the level of this factor, this can mean a difference of (over) 2 points on a scale from 1-10. As such, in this first term, both students and teachers have no proper prior sense of how many points they should obtain, or assign, as to (just) make the cutoff required for graduation. In the conversion formula, this factor is denoted by N, and we will often refer to it as the N-factor throughout the paper.

This all changes, however, in the second term (i.e. when students decide to take up the opportunity to retake one exam). In particular, the same conversion formula (N-factor), applies to the exam in the second term. As such, students and teachers know exactly how many points are required as to end up with an arithmetic average required for graduation. In this second term of students retaking exams, both the stakes and information surrounding the

grading process have increased considerably. In contrast to the first attempt, performance on this single test is often crucial for a student to graduate and the teacher knows exactly the conversion that will be applied as to generate grades from scores. This paper examines whether this potential for teacher discretion in grading is observed and what mechanisms and incentives might underlie this phenomenon.

Administrative secondary education data for the Netherlands, for the period 2007-2014. For each student, a list of background characteristics is known, together with the results on school examinations and national exams (for all subjects and both terms). We first formally test whether the student test score distribution experiences density discontinuities. We then disaggregate results by whether or not the student-subject score was obtained in the first or second term. Also we examine to what extent variation in such discontinuities occurs across or within schools. Next, we relate the size of the discontinuity to student- and school characteristics.

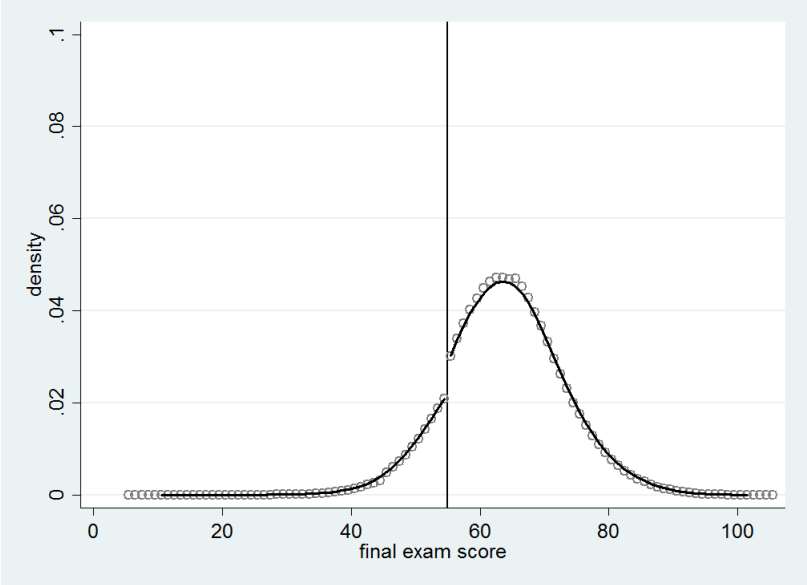
In order to explain the mechanisms behind these findings, we relate variation in school-level discontinuities over time to whether or not a school was under the scrutiny of more intense supervision by the Education Inspectorate. At the student-level we evaluate whether the Dutch context corroborates the "bad test day"-effect found in Sweden (Diamond and Persson, 2016) and the US (Dee et al., 2016), by linking the probability of observing a score just passed the cutoff to whether the performance on the national exam was lower than the school examination average. One specific artefact of the system is that each year, for some subjects, there is an unexpected change in the N-factor used for making the conversion from points to grade in the second term. This occurs if there were irregularities in the retake exam (e.g. an error in the question). These alterations, if implemented, will always result in a more lenient grading conversion and thus can only boost the student's grade. The consequence is that for these subjects, in contrast, teachers did not know the true conversion factor upon grading the second-term exams. As a matter of fact, depending on the size of the alteration, the actual student grade will be somewhat higher than expected based on the points assigned. We explore whether changes in the position of this discontinuity for these subjects are in accordance with this notion. Furthermore, we relate results to the nature of test (e.g. relatively many open-response or essay questions, as opposed to multiple-choice questions) as to gain further insight in potential mechanisms and incentives at play.

3 Preliminary Results

Figure 1 presents the overall discontinuity in test scores, based on all students and subjects for the period 2009-2011. The discontinuity is most notable just above the cutoff required to pass a particular subject. Disaggregating results by whether the score is obtained in the first or second term reveals that this discontinuity is completely driven by students who take up the opportunity to retake the exam for one subject.[1] Results show that the student-body population of the retry exam overwhelmingly consists of those who failed to graduate based on results after the first term, thereby underscoring the high-stake nature of the retake exam. Furthermore, there is considerable systematic variation in the size of this discontinuity, both across and within schools, and with estimates related to both student-, and subject characteristics. We find evidence for a "bad test day"-effect in the Netherlands, in that the likelihood of just passing the cutoff for passing a subject (e.g. 5.5) is positively related to whether a student's performance on the national exam is lower than that on the school examination. For subjects that underwent an unexpected change in the conversion formula in the second term (e.g. due to an error in a question), leading to a higher grade for all students, we find that the position of the discontinuity changes in accordance with the size of this alteration.

[1] We note that we have performed a series of manipulation test (McCrary tests) by which we formally determine whether there is evidence of a discontinuity in the density of the final exam score around the known passing cutoff (5.5).

Figure 1: Student test score distribution (preliminary)



References

Burgess, Simon and Ellen Greaves (2013), 'Test scores, subjective assessment, and stereotyping of ethnic minorities', *Journal of Labor Economics* **31**(3), 535–576.

Dee, Thomas S, Will Dobbie, Brian A Jacob and Jonah Rockoff (2016), The causes and consequences of test score manipulation: Evidence from the new york regents examinations, Technical report, National Bureau of Economic Research.

Diamond, Rebecca and Petra Persson (2016), The long-term consequences of teacher discretion in grading of high-stakes tests, Technical report, National Bureau of Economic Research.

Hanna, Rema N and Leigh L Linden (2012), 'Discrimination in grading', *American Economic Journal: Economic Policy* **4**(4), 146–168.

Jacob, Brian A (2005), 'Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools', *Journal of public Economics* **89**(5), 761–796.

Lavy, Victor (2008), 'Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment', *Journal of public Economics* **92**(10), 2083–2105.