Learning from Performance Information

Simon Calmar Andersen

Helena Skyt Nielsen

Abstract

Years of research on performance management has generally concluded that performance information is seldom used purposefully by public managers, and that it does not improve performance as intended. More recently both theoretical and empirical work has begun to focus on situations in which performance management may fulfil its promises. In a study of student testing in a low-accountability performance system in Denmark we examine the effect of a key component in education performance management, namely measuring student performance and giving performance feedback to teachers, parents and students so they can learn from the test results. Using as-good-as-random variation in students exposed to test due to a technical break-down in the IT system, we identify the effect of testing on student learning measured two years after the breakdown. Results show positive and statistically significant effects of about 0.1 standard deviations, which is comparable to much more expensive interventions. Implications and limitations in terms of increasing the level of student testing is discussed.

Introduction

Performance management is the systematic combination of setting organizational targets, measuring performance and evaluating performance based on these targets (Andersen 2008; Moynihan 2008). It has been a central feature of New Public Management reforms that have spread across most countries during the last thirty years (OECD 2007; Pollitt and Bouckart 2011). Disappointingly, reviews of existing research show that performance information is often not used (Kroll 2015) and average effects of performance management systems are small and uncertain due to lack of strong empirical designs (Gerrish 2016; see also Heinrich and Marschke 2010). Indeed, a number of studies in educational settings have shown how high-accountability pressures have caused both gaming (e.g. Deming et al., 2016; Jacob, 2005) and cheating (Jacob and Levitt, 2003) by teachers. This has spawned a renewed interest in the conditions under which the promises of performance management can be achieved. Empirically, Holm (2018) shows how managers in a system with learning forums and repeated performance measures actually use performance information to prioritize low-performing areas and that these prioritizations result in improved performance over time. Theoretically, Jakobsen and colleagues (2018) argue that it is crucial to find the right balance between the need for external accountability and internal learning in the organization.

In this study, we focus on one key component in performance management systems: measuring performance and using the information to learn how to improve. We argue that in an educational system with low accountability, testing students and feeding the results back to teachers, students, and parents may increase student learning for several reasons. Students may learn from taking the test (Roediger et al., 2011) and receiving constructive feedback (Black and Wiliam, 1998; Hattie and Timperley, 2007), teachers may use the test information to tailor teaching to the students' level (Banerjee et al., 2007), and the feedback may encourage parents to engage in

coproduction, i.e. contributing together with teachers in the "production" of their children's skills (Jakobsen and Andersen, 2013).

Identifying the effects of performance management systems has proven difficult and often rely on observational methods conditioning on observed control variables (Kroll 2015, Gerrish 2016). Isolating the effect of testing students and providing performance feedback, as we aim to do here, is even more difficult because there is often no variation in whether students are tested or not. Furthermore, the marginal effect of testing on the individual students should be expected to be small relative to the baseline effect of going to school and receive many hundred hours of teaching a year. Estimating small effects with adequate statistical power requires very large sample sizes. We therefore exploit a technical breakdown during the introduction of a nation-wide test system in Denmark involving 136,887 students. The break-down meant that it was as-good-as-random which students were less likely to take the test and thereby their teachers, parents and themselves would not get any test results. The test results are, by law, confidential and cannot be used publicly for comparing schools' performance. The Danish student testing system thereby is a good case for testing the effect of the key component in performance management systems in a low-accountability context. Using the break down as instrument for taking the test, we show that students who did not take the test because of the breakdown had lower performance in tests two years later. To further isolate the effect of individual feedback from effect on school managers and others, we use a fixed effects model to show that some of the effect can be ascribed to variation in student testing within schools.

In the next section we review theory and existing evidence for the effect of performance management systems. Then we lay out the study design before results and a number of robustness checks are presented. In the conclusion we discuss the implications of the results. Despite the positive effects identified, we do not claim that more tests would always further improve student

testing. Features of the Danish context, including the low-accountability and the low use of student testing prior to the introduction of the national test system, may be crucial contextual conditions for the observed effects.

Theory and Existing Evidence

Limitations of High-Stakes Accountability Systems

Performance management is based on a cyclic understanding of the management process in which performance targets are set, indicators chosen, performance is measured and evaluate against the targets, which may lead to adjustments of the performance targets and the cycle can start over (Andersen 2008; Moynihan 2008). In educational settings, which is the case we study, school accountability implies the use of administrative data to analyze performance and subsequently explicitly or implicitly reward or sanction schools with the purpose of improving student performance. Examples of explicit rewards/sanctions are withdrawal of autonomy and restructuring or closure, while an example of an implicit reward/sanction is information disclosure. Implicit rewards/sanctions are known to influence both housing prices and financial support, suggesting that they may in fact be as effective as explicit rewards/sanctions (Black, 1999; Figlio and Lucas, 2004; Figlio and Kenny, 2009).

The performance management cycle combined with a high level of accountability may seem to be an effective model for improving performance. Empirical research, however, has shown that on average, performance management systems do not improve performance. Gerrish (2016) conducted a meta-analysis of performance management studies and found very weak relationship between this practice and outcomes. Separate analyses for studies within education leads to the similar conclusion. Heinrich and Marschke (2010) reviewed the performance management within a principal-agent framework and point out how difficult it is to design a performance management

system that holds agents accountable for their performance, without creating strong incentives for gaming or cheating rather than learning – especially in a dynamic context in which the agents may anticipate and react to principals' attempts to adjust for unintended effects.

High-accountability systems within education has produced ambiguous results. According to a review by Figlio and Loeb (2011), evaluations of No Child Left Behind, high-stakes school accountability systems in the US, indicate that it improved student test performance, particularly in math, while evaluations of state-based or district-based systems find that the results are far more mixed. The magnitude of the estimated effects ranges from zero to about 0.30 of a standard deviation (SD), but a non-negligible part of the estimated effects is driven by actions that artificially improve school performance. Specifically, some studies document that schools invest in students, grades or subjects, which, in turn, contributes to improving the accountability rating (e.g. Deming et al., 2016; Chakrabarti, 2014; Krieg, 2011; Figlio and Rouse, 2006; Neal and Schanzenbach, 2010; Reback, 2008), while other studies show that students may be reclassified into special education or that schools invest in test-specific skills or efforts (e.g. Jacob, 2005). Jacob and Levitt (2003) find indications of outright cheating with student test results. Such strategic responses are difficult to manage in a dynamic setting where agents change behavior as they become familiar with the mechanisms of the accountability system (Heinrich and Marschke, 2010).

Some studies find slightly more promising effects of accountability pressure on student achievement. Rouse, Hannaway, Goldharber, and Figlio (2013) provide evidence that schools given lowest grading significantly change their instructional policies and practices, these responses substantially explaining subsequent test score gains. Using a comparative, interrupted time series approach based on all U.S. states, Dee and Jacob (2011) find significant effects of accountability on math but no robust effect for reading. However, their advocacy for the system is not without hesitation because, although effects are statistically significant for math, 60% of 4th graders

nevertheless fall below national proficiency standards and there is still no robust effect on reading proficiency. Based on a study of the impact of the National Assembly abolishing national testing in Wales, Burgess et al. (2013), on the other hand, strongly advocate test-based accountability.

In sum, high-stakes systems apparently produce positive effects, but they seem to be driven, to some extent, by strategic responses, e.g., on high-stakes testing and for students whose scores have the greatest consequences for school accountability.

The Potential for Learning in Low-Stakes Systems

These examples of failed performance management systems have given rise to renewed theoretical interest in how the seemingly contradictory purposes of accountability and learning may be unified. Jakobsen and colleauges (2018) argue that an internal learning regime in which professionals, among other things, are involved in the interpretation of performance goals, will increase motivation and ultimately performance. Instead of relying on political stakeholders to decide whether performance was above or below target, internal learning regimes delegate the interpretation of performance outcomes to the professionals. In a similar vein, Andersen (2005) argued that the politico-administrative system's accountability concerns should be separated from attempts to increase reflection and learning in the educational system, if political gaming and blame-avoidance logics should be avoided.

In line with these theoretical accounts, we suggest that there are reasons to believe that one key component in performance management systems, i.e., measuring performance and providing feedback to key stakeholders, may have positive learning effects without the strong incentives of high-stakes systems. More specifically, we argue that within education, the service area we study here, testing students and giving feedback to both students, parents and teachers may hold the potential for improving student performance.

As described in detail by Roediger et al. (2011), laboratory experiments show that children learn from taking tests. Children who take a test remember the content better one week later than children who repeat the material. Yet, it is not evident that these lab results translate into long-term effects in a real school context (Roediger et al., 2011). Outside the lab, students may also learn from getting teacher feedback based on the (Hattie and Timperley, 2007; Hattie, 2009). If standardized testing is used for formative assessment (and not just a summative assessment) where student performance is compared to a reference level, and test results constitute feedback in the sense of taking actions to alter the gap, there is ample evidence that student learning may improve (Black and Wiliam, 1998).

Parents may also react to the performance information entailed in the test results. Asking teachers to give students' test results to parents may increase parents' awareness that education is a prime example of coproduction in which inputs from both teachers and parents may help to improve student learning. Randomized controlled trials have shown that children whose parents were encouraged by schools to read and talk to their children did improve their language and reading skills (Jakobsen and Andersen 2013; Andersen and Nielsen 2016).

Finally, teachers themselves may learn from the test results. Randomized controlled trials testing interventions that use test results to target teaching to students' skill level have shown very positive results (Banerjee et al., 2007). Without tests of students it may be more difficult for teachers to assess students' skill levels and therefore to tailor the instruction to the individual students. Research on the accuracy of teacher expectations has shown that teachers tend to be downward biased in their evaluation of the learning potential of ethnic minority students (for reviews, see Jussim and Harber, 2005; Tenenbaum and Ruck, 2007).

We are not testing each of these potential channels (students, parents, teachers) separately, but we use them to substantiate our main expectation *that introducing standardized student testing*

and requiring that teachers convey test results to students and parents—but without implicit or explicit high-stakes such as publication of results or sanctions/rewards based on performance results—will increase student learning.

Besides testing the main expectation, the data allow us to test a number of supplementary research questions. First, it is important to know if effects are different for different groups of students. Opponents of nationwide compulsory testing fear that the weakest students may suffer from compulsory testing (see Deming et al., 2016).

Secondly, some of the effect of performance measurement and feedback may arise without the interference of the manager of the organization. This is, obviously, not because managers would not matter for performance management, but because some of the learning that may take place in the organization, may be generated by reflection processes that do not directly involve the manager and leave more autonomy to the professionals (cf. Jakobsen et al. 2018). By comparing students tested and not-tested within schools, we are able to separate out any general effects of school managers and other school characteristics. School managers might still have an effect if they discuss individual students results with the teachers, but the analysis gives some idea of how much of the effect is related to school-invariant factors.

Thirdly, schools with low-SES students and low exit exam grades would be in more need for interventions that could help them improve their performance. But even in a low-stakes system they may be more reluctant to comply with a new test regime. We examine on the one hand whether schools with high shares of disadvantaged students oppose the new performance management system by not signing their students up for the test system—and on the other hand whether these schools benefit more from taking the tests.

We do not want to overstate the claim that testing increases learning. The specific effects of such a performance management system depends on the details such as the specific

accountability procedures and the prior use of tests. We therefore describe the institutional context in some detail in the next section and use this in our concluding discussion of the implications and limitations of the study.

Methods

Test-Based Accountability in the Low-Stakes Danish System

Until 2010, the performance of Danish students was not systematically evaluated until 8th grade (approximately age 15). However, based on poor Programme for International Student Assessment (PISA) results in 2000 and 2003, a subsequent OECD report (2004), and recommendations from various national committees, the Danish Parliament opted for a cultural shift that made the assessment of learning an integrated part of schooling and implemented systematic, standardized compulsory evaluation of student performance in all primary and lower secondary public schools.

At the beginning of 2006, a nationwide school accountability system was approved to ensure continual quality assessment and quality improvement of the Danish public schools.[1] The initial accountability system comprised nationwide testing of students ten times from 2nd to 8th grade, annually updated individual plans for students, compulsory 9th grade exit exams, and annual quality assessment reports at the level of local authorities. The assumption was that these steps would provide students, parents, teachers, school principals, and local authorities with information about the input necessary from schools to ensure continual quality development.[2]

The policy context is best described as having limited accountability. In an international comparison, Danish school accountability is, as yet, based on low-powered incentives. No explicit

[1] Detailed in the Public School Law (Law no. 313, April 19, 2006 and Law no. 572, June 9, 2006) and described by the Danish Ministry for Children, Education and Gender Equality (2011a). Within the framework of the national law, public schools in Denmark are governed by and accountable to local authorities.
[2] Danish Ministry for Children, Education and Gender Equality (2011a).

proficiency standard has been set for local authorities, schools, subgroups or individuals.[3] However, each school is endowed with an annual socio-economic index based on sex, ethnicity, and parental socio-economic status (SES), enabling schools to compare test results with the national average at schools with a similar socio-economic index. Schools are told what the gap is between the national average and their test score for each of the ten tests and given an indication of whether the gap is statistically significant or not.

An essential element of the Danish test-based accountability system is the compilation of the annual quality assessment report. Every local authority is obligated to produce, discuss, and publish a quality assessment report based on input from its schools. The report particularly concerns academic progression over time and potentially significant deviance from other schools with similar socio-economic indices but also measures taken to deal with unsatisfactory results. Poor academic performance does not result in automatic repercussions, but 15% of students were affected by school consolidations, determined to some extent based on performance, during the 2010-2011 and 2011-2012 school years (Beuchert, Humlum, Nielsen, and Smith, 2018).

The quality assessment report incorporates information about the school and average local authority test score on the national tests, in addition to other measures of academic performance and academic progression, such as results from the 9th grade exit exams and other tests, participation rates in the tests, and individual circumstances such as the percentage of students with special needs. The evaluation of academic performance forms the basis for the quality assessment and the future objectives of the school and the local authorities as a whole. However, the average test score of schools is confidential (no explicit or implicit rankings are allowed), which means the public

[3] The Public School Law of June 7, 2013 introduced proficiency standards and set a national target of 80% proficiency in reading and math.

version of the report is not allowed to reveal results for schools individually,[4] preventing parents and the media from using test results for selecting schools or for naming and shaming.

Nationwide testing was initially legislated in Denmark when the nationwide school accountability system was introduced in 2006. However, due to technical and other challenges, testing was postponed until the 2009-2010 school year. Beginning that year, ten compulsory standardized national tests were introduced in the Danish public schools from 2nd to 8th grade. Students were tested on various subjects throughout their school career, though the main emphasis is on reading and math.[5]

The compulsory tests take place from January to April at the end of grades 2 to 8, as summarized in Table 1. In addition to the compulsory test, students have the option of taking the test voluntarily twice. Questions in the optional tests are drawn from the exact same pool of questions, though students rarely encounter the same question twice because of the size of the pool and because the tests are adaptive (we return to this below). Teachers register students to take the optional tests in the autumn at the grade level of the compulsory test, or in the autumn at the grade level before or after the compulsory test.

TABLE 1
COMPULSORY NATIONAL TESTS

	Grade						
Subject	2nd	3rd	4th	5th	6th	7th	8th
Reading	X		X		X		X
Math		X			X		
English						X	
Geography							X
Physics/Chemistry							X
Biology							X

[4] Danish Ministry for Children, Education and Gender Equality (2011b).
[5] See Beuchert and Nandrup (2018) for a detailed description of Danish national tests.

Schools and teachers are required to make sure students take the tests during the test period, but they are free to decide when to book tests during the test period. Some local authorities urge schools to administer the test during a narrow time window to gain precise measures of yearly performance progression.[6] Booking and rebooking tests opens from one week prior to the test period until the end of the test period. Either the teacher or the school secretary can be responsible for booking the test, and the class can be divided in two for testing. The testing system is used more intensively in March and April than in January and February (see Figure A1 in Appendix A).

The tests are designed to test proficiency in three different domains, or profile areas, in each subject. For example, the reading test focuses on language comprehension, decoding, and reading comprehension, while the math test centers on numbers and algebra, geometry, and applied mathematics. Thus, the national tests in reading and math are by no means exhaustive, although they do cover a major swath of what is considered testable content and crucial learning goals.

The national tests are IT-based and self-scoring, which means teachers are not involved in grading them. Our results are therefore not driven by subjective teacher gradings.

The tests are adaptive, which means they adapt to the child's abilities during the test, challenging children according their skill level. The test starts out with a moderately difficult question. If the question is answered correctly, the student gets a more difficult question, but if the question is answered incorrectly, the student gets an easier question, and so on. When the test results for all three domains reach a sufficiently statistical certainty, the test ends.

The test scales measure absolute ability within each domain. Approximately 183,000 individual test results, or 15,000-21,000 results for each of the ten national tests, were used to score the tests when they were introduced in 2010 (Beuchert and Nandrup, 2018). These results set the

[6] Data on the timing of tests shows that students booked early in the test window in 2010 are also more likely to be tested early in the subsequent test in the same subject.

norm for the scale. The test is revised annually to eliminate items that are deemed erroneous or noisy.

The underlying psychometric model is a Rasch model (see e.g. Bond and Fox, 2007), with test results for each domain measured on a Rasch scale from -7 to 7, seven representing the highest skill level. The score for a domain is thought to measure student ability for this exact domain. The teacher receives a detailed report of the student answers as well as a one-page summary to be shared with the student and the parents. The summary includes the overall score on a five-point scale as well as the score for each of the three test domains on a similar five-point scale.[7] The teacher is required to provide individual feedback to each student when they share the summary. In sum, the test provides students and teachers with relatively precise information on each student's ability on different domains within a specific subject.

Identifying Variation: A Major Technical Breakdown

To estimate the effect of testing on student learning, let Y_i^{j+d} be a measure of student achievement for individual *i* at grade *j+d;* let T_i^j indicate whether the student was exposed to nationwide testing in grade *j*, *d* years prior to measuring the outcome; and let X_i include relevant control variables all measured at age seven plus indicator variables for grade levels. We model the relationship between student achievement and test taking as:

$$Y_i^{j+d} = X_i\beta + \alpha_i T_i^j + \varepsilon_i \quad (1)$$

[7] This scale resembles the grading scale with the following approximate distribution of scores: 10% 'substantially below the mean'; 25% 'below the mean'; 30% 'at the mean'; 25% 'above the mean'; and 10% 'substantially above the mean'.

Compulsory nationwide testing was introduced universally, which means there is no obvious way to measure the impact of test taking, *T*, on student achievement. Even though the tests were mandatory, not all students took them as some were exempted, while others—school managers, teachers, parents or students—were uncooperative. This variation in test taking behavior may be correlated with ε_i if, for example, unobserved variables related to the student population influence test taking decisions. Also, *T* may be correlated with α if test taking is based on expected gains.

To solve this, we exploit a major crash in the IT system during the first year of the system. The IT system proved to be exceedingly vulnerable to, for instance, a large number of simultaneous users. As a consequence, the system was rather unstable in the beginning of the test period in 2010, and in March 2010 the system crashed, which led to its closure for almost two weeks. The crash meant that all students booked for the test in the periode March 2-12, 2010 were unexpectedly exempted from taking the compulsory test (see Table 2). We know which students were booked for the aforementioned period, but disregard test results from March 1-2, 2010 and March 11-12, 2010 since accurate information about who the IT problems affected is unavailable.

TABLE 2
PERFORMANCE OF THE NATIONAL TEST IT SYSTEM IN 2010

Test Period	**System**	**Test Taking**
January 20 – March 1, 2010	Open	Compulsory
March 2 – March 10, 2010	Closed	Retake voluntary
March 11– March 12, 2010	Open	Voluntary
March 15 – April 29, 2010	Open	Compulsory

We exploit the fact that the crash—and the sudden exemption from taking the otherwise compulsory test—was unexpected for the students and the teachers. Thus, teachers and students

exposed to the crash were no less prepared for the test compared to the rest of the population. Some teachers rebooked some of the students that were affected by the crash, but the crash made it less likely that an affected student ended up taking the test. We therefore pursue an IV strategy that uses an indicator for being exposed to the IT breakdown as an instrumental variable for taking the test, *T*. We use two-stage least squares (2SLS) to estimate the parameters of interest. The IV analysis identifies the effect of taking the test for the "compliers", i.e., students who are not being tested if they are unexpectedly exempted from taking a compulsory test. We define the instrumental variable as an indicator variable for whether the student was booked during the crash or not.

Threats to Validity

There are two channels for potential non-random selection of student testing that could threat the validity of the instrument: Selection in booking of test sessions and selection in exposure to the crash.

Non- random registration of bookings during the crash. As described above, the teacher books the test session in advance for a specific date and time. Figure A1 in the appendix illustrates that test behavior follows a smooth pattern in 2010, 2011 and 2012, where an ever-increasing number of students are tested over the period, with the greatest number of tests taken in March and April. In Figure A2, we have reconstructed the missing information by filling in the available information about test behavior during the crash period. Even after this reconstruction, it is quite evident that a substantial number of observations are missing during the crash period. The number of registered bookings during the crash period does not appear to compare to what one would expect from simple extrapolation of test behavior in the adjacent periods. This is most likely a natural consequence of the IT problems. If the missing observations are random, it is not necessarily a problem for our empirical strategy. In Figures A3-A5, we examine the missing

observations for eight of the tests in more detail. It appears that the booking information during the crash period is, practically speaking, complete in Figure A3 for the reading test in 2nd and 4th grade, and in Figure A4 for the math test in 3rd grade. However, Figure A5 clearly shows that the information for 8th grade is incomplete as around 80% of the observations are missing during the crash period (if we assume that the test activity would be smooth without the crash). As a result, we only use booking information for the early reading tests (2nd-6th grade) and for the early math test (3rd grade). As Figure A2 shows, information is only available for completed tests for March 11-12, 2010. Hence, information from those two dates comprises only voluntary testing and therefore we disregard completed tests from those two dates. Furthermore, some of the tests on March 1-2, 2010 are recorded as ordinary completed tests and some are recorded as booked during the crash as well as completed, which is why tests from those two dates are also disregarded. Consequently, the employed instrument measures whether students are booked for a test from March 3-10, 2010.

Non-random exposure to the crash. Students exposed to the crash were booked in the first part of the test period, which may reflect the fact that their teachers planned to use test results formatively, at least more so than teachers booking the test at the end of the period. As a robustnesss check, we investigate whether the timing of the test matters by narrowing the analysis to students booked for the test +/- two weeks around the crash (i.e. before April 1, 2010).

It may also be the case that students exposed to the crash were tested later in the period the next time, for instance, because teachers would then avoid being exposed to an unstable test system again. If, on average, students exposed to the crash were tested later in the school year the next time, they would have more time to learn from the teaching. Although the tests are supposed to measure annual progress, the test window amounts to almost one-third of the school year, which could bias the results (see e.g. Fitzpatrick et al., 2011). As an additional robustnesss check, we investigate whether the timing of the subsequent test matters for the results.

Data and Sample

We use register-based data from Danish registries for children born from 1996-2002. This dataset includes all students in 2nd to 6th grade for the 2009-2010 school year who were no more than one year ahead of or behind the schedule. Our goal is to investigate the effect of taking a test in 2010 on future test results. Due to the testing schedule described in Table 1, we are able to investigate the effect of taking the reading test in 2010 on reading scores in 2012, as well as the effect of taking the math test in 2010 on math scores in 2013. We standardize the outcome variables to have mean of zero and standard deviation of one. We also study the standardized scores for each of the three test domains separately

We sample individuals with a reading score in 2012 and individuals with a math score in 2013.[8] Table 3 shows that in these two samples more than 75 % of the students were unaffected by the crash and did take the test as required. 5-7% were exposed to the crash and did not take the test. However, around 7% took the test despite being exposed to the crash, and around 10% did not take the test even though they were not exposed to the crash. The main explanation for a missing test score is uncooperativeness, in the sense that students, parents, teachers, and headmasters decided not to obey the stipulated law requiring students to take the test.[9] We expect the extent of this type of behavior to be substantial in light of the public dispute about the potential benefits or harms of standardized testing.[10] These non-compliers is the reason behind our IV design.

[8] These test scores are our outcomes of main interest. In principle, "not being tested" in reading in 2012 or in math in 2013 could also be studied as alternative outcomes. However, these outcomes do not vary much: 91.8% of those exposed to a crash are tested in reading in 2012 compared to 91.0% of those not exposed to a crash.
[9] A few of these individuals could be returning from abroad or from private schools (<1%).
[10] For instance, the teachers' union represents a critical voice in this debate (see folkeskolen.dk).

TABLE 3

STUDENTS TESTED IN READING 2012 AND MATH 2013

Took the test 2010	Exposed to Crash 2010	
	Yes	No
Yes	Reading: 7.1% Math: 6.7%	Reading: 76.1% Math: 77,5%
No	Reading: 7.2% Math: 5.1%	Reading: 9.6% Math: 10.7%

Note: $N_{reading}$=151,375; N_{math}=51,880

For the subsequent empirical analysis, we select all individuals who took the compulsory reading or math test in 2010 or who were exposed to the crash, but did not take the test. This amounts to 136,887 students in reading and 46,338 in math. We exclude students who were neither exposed to the crash or took the test, and our empirical results should be interpreted as *conditional on* taking the test or being exposed to the crash (or both).

We collect a rich set of background characteristics consisting of demographic variables, education, and labor market status of the parents. All variables are measured at age seven, which is at school entry and therefore not affected by subsequent test taking. In some of the analyses, we distinguish between high and low SES, which is defined as having at least one parent who completed a college education and no parents with a college education, respectively.

TABLE 4

REGRESSION OF CRASH ON BACKGROUND CHARACTERISTICS

	Booked during Crash (2010)					
	Reading			Math		
	2nd, 4th or 6th Grade			3rd Grade		
	Coefficient	Standard error	Mean	Coefficient	Standard error	Mean
Child Characteristics						
Boy	-0.00136	(0.00213)	0.502	-0.000402	(0.00309)	0.502
Young for Grade	-0.0139	(0.00895)	0.022	0.0222	(0.0140)	0.023
Old for Grade	0.00135	(0.00336)	0.177	0.00250	(0.00488)	0.171
Non-Western Immigrant or Descendant	-0.0117	(0.00849)	0.058	-0.0147	(0.0125)	0.056
Western Immigrant or Descendant	0.00779	(0.00889)	0.038	0.0224	(0.0137)	0.038
Only Child	0.00599	(0.00473)	0.102	0.00399	(0.00726)	0.100
No. of Children in Family	-0.00419*	(0.00242)	2.323	-0.00679**	(0.00346)	2.325
Youngest Child	0.00271	(0.00284)	0.302	0.00699	(0.00442)	0.300
Middle Child	0.00647*	(0.00345)	0.283	0.00432	(0.00507)	0.284
Other	0.00607	(0.0100)	0.021	0.0240	(0.0151)	0.021
Parental education						
Mother, High School	0.00466	(0.00500)	0.071	0.00355	(0.00736)	0.069
Mother, Vocational Education	0.00194	(0.00343)	0.376	0.000515	(0.00544)	0.371
Mother, Two-year College	0.000923	(0.00605)	0.045	-0.00177	(0.00904)	0.047
Mother, Four-year College	0.000136	(0.00459)	0.224	-0.00525	(0.00662)	0.233
Mother, M.Sc. or Ph.D. Degree	0.0129*	(0.00747)	0.077	-0.0100	(0.00978)	0.083
Father, High School	-5.05e-05	(0.00615)	0.051	0.00339	(0.00941)	0.050
Father, Vocational Education	-0.00176	(0.00320)	0.415	-0.00724	(0.00466)	0.414
Father, Two-year College	-0.0111**	(0.00492)	0.068	0.00639	(0.00767)	0.070
Father, Four-year College	0.00250	(0.00548)	0.120	0.000450	(0.00740)	0.124
Father, M.Sc. or Ph.D. Degree	0.00262	(0.00810)	0.097	0.00670	(0.0102)	0.102
Parental Labor Market Status and Income						
Mother, Self-Employed	-0.000122	(0.00612)	0.032	0.0208*	(0.0101)	0.032
Mother, Enrolled in Education	0.00631	(0.0109)	0.010	0.0524***	(0.0196)	0.010
Mother, Unemployed	0.00553	(0.00653)	0.033	0.00150	(0.0126)	0.024
Mother, Other Activity	-0.000265	(0.00380)	0.138	-0.00553	(0.00653)	0.126
Father, Self-employed	-0.00312	(0.00431)	0.083	0.00371	(0.00640)	0.081
Father, Enrolled in Education	0.00904	(0.0240)	0.002	0.0659	(0.0507)	0.001
Father, Unemployed	-0.00205	(0.00825)	0.021	-0.0130	(0.0147)	0.015
Father, Other Activity	0.000363	(0.00439)	0.070	0.00895	(0.00693)	0.063
Income DKK 125,000-220,000	0.00282	(0.00517)	0.230	0.0120	(0.00806)	0.206
Income DKK 220,000-345,000	0.00435	(0.00550)	0.430	0.00736	(0.00888)	0.415
Income > DKK 345,000	0.0108	(0.00719)	0.261	0.00784	(0.0108)	0.303
Mother, Unemployed 1-50%	-0.00156	(0.00311)	0.173	0.00603	(0.00463)	0.188
Mother, Unemployed >50%	0.00625	(0.00687)	0.032	0.00705	(0.0114)	0.025
Father, Unemployed 1-50%	0.00175	(0.00412)	0.088	0.00855	(0.00669)	0.079
Father, Unemployed >50%	-0.00421	(0.00880)	0.019	-0.0115	(0.0153)	0.014
Number of Observations	136,887			46,338		

Note: Constant terms, grade-level fixed-effects and dummies for missing values are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

Summary statistics are given in Table 4, which shows the results from a regression of being exposed to the crash on observable characteristics for the estimation sample. Most parameter estimates are small in magnitude and only a few are statistically significant.[11] In the empirical analyses, we show that results are robust when control variables are added. Furthermore, we perform the analyses separately by subgroups.

Results

Main effects for all students

Table 5, column 1 shows the main results for reading for all students. The first-stage results reflect that being exposed to the crash reduces the probability of being tested by 50%. The reduced-form results indicate that being exposed to the crash is associated with a reduced reading score of almost 5% of a SD. The second-stage results show that taking a reading test increases the test score as measured two years later by about 9% of a SD.

The magnitude is similar to what has been found in the U.S. Figlio and Ladd (2008) find an effect of 0.08 SD and Burgess et al. (2013) find an effect of 0.07 SD. On average, student progress in reading for an entire school year is 32-40% of a SD during the relevant grades (Lipsey et al., 2012).[12] Thus, taking the test exacerbates the two-year learning effect by 12.5%.[13] We cannot separate how much of the effect is due to the effect of students taking the test and teachers receiving test results and giving feedback to students and parents.

[11]The pattern is similar when we run a school-level regression, and the parameter estimates are similar when we run the regression for students booked before April 1st. In the empirical analyses, we present robustness analyses where we study a narrow test window around the crash.

[12] The numbers measure student progression from spring to spring and decline with grade. From grades 3-4, 4-5 and 5-6, the student progression is calculated to be 36, 40 and 32% of a SD in reading and 52, 56 and 41% in math.

[13] This is 9/(32+40) = 0.125.

Table A1 in Appendix A shows that the results are almost unaffected when controls are added and that the results do not vary much across profile areas. This supports the assumption that exposure to the test was as-good-as-random.

TABLE 5

THE EFFECT OF TAKING A READING TEST IN 2ND, 4TH, AND 6TH GRADE ON READING PERFORMANCE TWO YEARS LATER

	(1) All	(2) Males	(3) Females	(4) Low SES	(5) High SES	(6) Non-Western Immigrants
First Stage: Tested						
Crash	-0.503***	-0.506***	-0.500***	-0.502***	-0.504***	-0.505***
	(0.0210)	(0.0214)	(0.0213)	(0.00199)	(0.00220)	(0.00610)
Relative to Overall First Stage		1.006	0.994	0.998	1.002	1.004
Reduced Form: Test Score						
Crash	-0.0462***	-0.0546***	-0.0380***	-0.0424***	-0.0503***	-0.0750**
	(0.0106)	(0.0135)	(0.0120)	(0.00921)	(0.00916)	(0.0296)
Second Stage: Test Score						
Tested	0.0918***	0.108***	0.0759***	0.0845***	0.0997***	0.147**
	(0.0218)	(0.0273)	(0.0245)	(0.0265)	(0.0253)	(0.0713)
Number of Observations	136,887	68,746	68,141	75,345	61,542	7,875

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

Table 6, column 1 similarly shows the results for math for the overall sample. The first-stage results indicate that being exposed to the crash reduces the probability of being tested by 43%; the retake probability is thus slightly higher than for reading. The point estimate in the second-stage results is 7%, which is similar to the point estimate for reading, though insignificant due to the smaller sample size. On average, student progression in math over those three school years is 150% of a SD (Lipsey et al., 2012). Thus, taking the test exacerbates the three-year learning effect by about 5%. Table A2 in Appendix A shows that the results are almost unaffected when controls are added and that they do not vary much across profile areas.

TABLE 6

THE EFFECT OF TAKING A MATH TEST IN 3RD GRADE

ON MATH PERFORMANCE THREE YEARS LATER

	(1) All	(2) Males	(3) Females	(4) Low SES	(5) High SES	(6) Non-Western Immigrants
First Stage: Tested						
Crash	-0.436***	-0.441***	-0.431***	-0.460***	-0.408***	-0.512***
	(0.0306)	(0.0319)	(0.0323)	(0.00328)	(0.00347)	(0.00998)
Relative to Overall First Stage		1.011	0.989	1.055	0.936	1.174
Reduced Form: Test Score						
Crash	-0.0303	-0.0273	-0.0322	-0.0472***	-0.00941	-0.0371
	(0.0216)	(0.0265)	(0.0253)	(0.0163)	(0.0188)	(0.0525)
Second Stage: Test Score						
Tested	0.0701	0.0624	0.0757	0.103**	0.0233	0.0717
	(0.0497)	(0.0586)	(0.0597)	(0.0484)	(0.0703)	(0.107)
Number of Observations	46,338	23,241	23,097	24,754	21,584	2,617

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

Heterogeneity across Groups of Students

The results above suggest that test exposure is positively related to subsequent test performance. Tables 5 and 6 (columns 2-6) show the results divided by gender, SES and immigrant status. We find no evidence of harmful effects for supposedly weak students. On the contrary, the point estimate of being tested in reading is higher for non-Western immigrants than for other students, and the point estimate of being tested in math is high and significant for students with low SES but close to zero for students with high SES. These results are evidence against the concerns that weaker students are harmed by testing (Deeming et al. 2016).

Separating out School-Invariant Factors such as School Managers

By including school fixed effects in the model, we essentially compare students within schools that either take or do not take the test because of the crash. Since they all have the same school manager, all general effects of school managers and other school characteristics are cancelled out of the model. As mentioned, school managers might still have an effect if they discuss individual students' results with the teachers, but the school fixed effects models leave out all school-invariant variation. Table 7 shows the results of the school fixed effects analyses for reading.[14] The estimates are roughly halved when school fixed effects are added, but they are still statistically significant. One exception is the effect for non-Western immigrants which doubles after inclusion of school fixed effects (while the sample of cooperative schools with non-Western immigrant pupils is halved).

We think it is reassuring that the effects are generally smaller but still statistically significant. This suggests that the effect is partly driven by school variation (such as school management and test culture), and partly driven by responses particular to the student's test experience or the subsequent student-teacher-parent interaction.

[14] Adding school fixed effects to the math estimations does not make much sense because only one cohort is included in our study compared to three cohorts for reading.

TABLE 7

THE EFFECT OF TAKING A READING TEST IN 2ND, 4TH, AND 6TH GRADE ON READING PERFORMANCE TWO YEARS LATER

SCHOOL FIXED EFFECTS

	(1) All	(2) Males	(3) Females	(4) Low SES	(5) High SES	(6) Non-Western Immigrants
First Stage: Tested						
Crash	-0.402***	-0.401***	-0.402***	-0.409***	-0.394***	-0.402***
	(0.00148)	(0.00210)	(0.00211)	(0.00201)	(0.00222)	(0.00637)
Relative to Overall First Stage		0.998	1.000	1.017	0.980	1.000
Reduced Form: Test Score						
Crash	-0.0261***	-0.0546***	-0.0380***	-0.0424***	-0.0503***	-0.0750**
	(0.00818)	(0.00965)	(0.00881)	(0.00921)	(0.00916)	(0.0296)
Second Stage: Test Score						
Tested	0.0517**	0.0509	0.0500*	0.0466	0.0538*	0.281***
	(0.0210)	(0.0311)	(0.0285)	(0.0290)	(0.0304)	(0.0950)
Number of Schools	1,850	1,748	1,698	1,785	1,666	991
Number of Observations	136,887	68,746	68,141	75,345	61,542	7,875

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

Schools with High Shares of Disadvantaged Students

To examine whether schools with high shares of disadvantaged students were less likely to comply with the new performance management system by not booking tests for their students, we regress an indicator for signing up for the compulsory test on a set of school characteristics. We find that schools with very low exit exams (as measured by 9th grade exams in the 2008-2009 school year), schools with a high proportion of low SES students, and schools with more than 15% immigrants tend to be less likely to sign up for the compulsory tests (i.e. more likely to be uncooperative). Table 8 shows the results.

TABLE 8

REGRESSION OF INDICATORS OF TEST BEHAVIOR ON SCHOOL CHARACTERISTIC

	Sign Up for Compulsory Test in 2010 (Sample: All)			
Reading 2nd, 4th and 6th Grade				
Lowest Decile Exit Exams		-0.0370***	0.00327	0.00870
		(0.0127)	(0.0149)	(0.0147)
Highest Decile Exit Exams		0.0208***	0.00966	0.0118
		(0.00766)	(0.00852)	(0.00852)
Missing Exit Exams		-0.0138**	-0.00692	-0.00627
		(0.00661)	(0.00647)	(0.00641)
Lowest Quarter SES	-0.0673***		-0.0669***	-0.0541***
	(0.00735)		(0.00742)	(0.00721)
Highest Quarter SES	-0.00733		-0.00913	-0.0103
	(0.00638)		(0.00688)	(0.00688)
More than 15% Immigrants	-0.0311***		-0.0320***	-0.0173
	(0.0115)		(0.0121)	(0.0116)
Constant Term	0.927***	0.907***	0.928***	0.919***
	(0.00381)	(0.00353)	(0.00416)	(0.0101)
Additional Controls				x
Mean of Dep. Variable		0.9043		
Number of Observations		151,375		
Math 3rd Grade				
Lowest Decile Exit Exams		-0.0252	0.0131	0.0191
		(0.0153)	(0.0208)	(0.0205)
Highest Decile Exit Exams		0.0187	0.0233	0.0313**
		(0.0145)	(0.0159)	(0.0159)
Missing Exit Exams		-0.150	-0.146	-0.134
		(0.118)	(0.114)	(0.100)
Lowest Quarter SES	-0.0756*		-0.0603**	-0.0430***
	(0.0397)		(0.0241)	(0.0135)
Highest Quarter SES	-0.0239**		-0.0357***	-0.0325***
	(0.00944)		(0.0126)	(0.0119)
More than 15% Immigrants	-0.0332**		-0.0518***	-0.0383**
	(0.0134)		(0.0193)	(0.0169)
Constant Term	0.922***	0.926***	0.955***	0.946***
	(0.0220)	(0.00459)	(0.00972)	(0.0140)
Additional Controls				x
Mean of Dep. Variable		0.8932		
Number of Observations		51,880		

Note: Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

In Tables 9 and 10, we examine whether the effects of testing differ for schools with high shares of disadvantaged students. First, we analyze the compliers to characterize the population who respond to the instrument and thus provide us with identifying variation. We divide the sample into subsamples and compute the ratio of the first-stage coefficient to the overall first-stage coefficient. Then, we interpret the second-stage estimates as impact estimates of taking each of the tests.

For the reading test, students at the schools in the lowest decile of the grade distribution are more likely to retake the test (i.e. they respond less to the crash), while the students at the schools in the highest decile of the grade distribution exercise their option to avoid taking the test (i.e. they respond more to the crash). The results in the second-stage regression reveal that the impact of being tested is indeed high for the students attending schools in the lowest decile, while it is literally zero for students attending schools in the highest decile. This may reflect that these schools already have sufficient good evaluation practices, even without compulsory nationwide testing, and therefore the compulsory tests make no difference in their case. The point estimates are not statistically different across subgroups. The pattern is not confirmed for the math test, where the results are generally less precisely estimated.

When we focus on schools at the bottom quarter and top quarter of the SES distribution, we see a similar tendency. Although there is no difference in the probability to comply, the impact of taking the test tends to be higher for students attending schools at the bottom of the SES distribution than those attending schools at the top of the distribution. This is true for reading as well as math.

Students at schools with many non-Western immigrants are more likely to retake the reading test (i.e. they are less affected by the instrument) and the impact of being tested in reading and math is high for this group. Importantly, this effect is conditional on the immigrant status of the individual, which the previous subsection also showed as important for the impact of being tested.

TABLE 9

HETEROGENEITY BY SCHOOL CHARACTERISTICS

THE EFFECT OF TAKING A READING TEST IN 2ND, 4TH, AND 6TH GRADE

ON READING PERFORMANCE TWO YEARS LATER

	(1)	(2)	(3)	(4)	(5)
	Lowest Decile Exit Exam	Highest Decile Exit Exam	Lowest Quarter SES	Highest Quarter SES	More than 15% Non-W. Immi.
First Stage: Tested					
Crash	-0.402***	-0.545***	-0.502***	-0.500***	-0.472***
	(0.00756)	(0.00630)	(0.00298)	(0.00295)	(0.00477)
Relative to Overall First Stage	0.799	1.083	0.998	0.994	0.938
Reduced Form: Test Score					
Crash	-0.0862**	0.00279	-0.0552***	-0.0415***	-0.0992***
	(0.0341)	(0.0218)	(0.0140)	(0.0114)	(0.0232)
Second Stage: Test Score					
Tested	0.214	-0.00511	0.110**	0.0829**	0.210***
	(0.136)	(0.0560)	(0.0467)	(0.0378)	(0.0811)
Number of Observations	5,222	8,024	33,840	35,160	12,779

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

TABLE 10

HETEROGENEITY BY SCHOOL CHARACTERISTICS:
THE EFFECT OF TAKING A MATH TEST IN 3RD GRADE
ON MATH PERFORMANCE THREE YEARS LATER

	(1)	(2)	(3)	(4)	(5)
	Lowest Decile Exit Exam	Highest Decile Exit Exam	Lowest Quarter SES	Highest Quarter SES	More than 15% Non-W. Immi.
First Stage: Tested					
Crash	-0.306***	-0.367***	-0.512***	-0.336***	-0.451***
	(0.0113)	(0.00969)	(0.00504)	(0.00419)	(0.00712)
Relative to Overall First Stage	0.702	0.842	1.174	0.771	1.034
Reduced Form: Test Score					
Crash	0.00572	-0.0564	-0.0325	0.000952	-0.105***
	(0.0526)	(0.0493)	(0.0245)	(0.0225)	(0.0403)
Second Stage: Test Score					
Tested	-0.0189	0.159	0.0638	-0.00286	0.231
	(0.302)	(0.231)	(0.0792)	(0.137)	(0.142)
Number of Observations	1,833	2,726	10,596	13,915	5,030

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

Robustness checks

Tables A3 and A4 in the appendix present robustness checks with respect to the timing of the tests. Column 1 reproduces the main results from tables 5 and 6, while column 2 shows the results for a narrow window (+/- two weeks) around the crash, where booking decisions are supposedly more random. The results are largely robust to narrowing the test window. Column 3 shows the results when we include a control variable for whether the subsequent test was taken late (April 1, 2010 or later) or not. The effects are robust and thus not driven by the timing of subsequent tests.

Conclusion

Evaluations of performance management systems tend to find no or very small effects on organizational performance. These findings have fostered a renewed interest in finding components of performance management systems and conditioning contexts that may foster some of the positive effects that has been hoped for. We do not study the effect of low or high accountability pressures, but within a system of relatively low-stakes accountability we test the effect of a core component of performance management systems, namely measuring performance and feeding this performance information back to the immediate users—in this case teachers, students and parents.

Our study suggests beneficial effects of testing the students of around .09 of a standard deviation. This is comparable to effect sizes of higher-powered accountability systems in the US but may come without the down-side of gaming and cheating. These effect sizes we find compare to the effect of reducing class size by 3-4 students (Heinesen 2010) or having a coteacher in the classroom most of a school year (Andersen et al. 2016), which are much more costly policies. The effects are positive across student background, and about half of the effect can be ascribed to within-school variation which excludes general effects of the school managers. Furthermore, we find that schools with low grades, many students with low socio-economic statuses, and many non-Western immigrants are less likely to comply with the new performance management system, but students at these schools tend to benefit more from being tested.

Having demonstrated these positive effects, some caveats are important to consider. The Danish nationwide standardized tests where introduced in a regime where students were not systematically tested or graded until 8th grade (age 15), before the tests were introduced. The system introduced 10 compulsory tests taken across seven years of schooling from 2nd grade to 8th grade. So, this is a low-dosis performance management system. Effects of the first few, standardized,

nationwide tests are may be larger than the effect of increasing the number tests in systems that already use systematic testing.

Also, even though we do not test the effect of high- versus low-accountability we would expect that effects may differ in high-accountability systems because of the stronger incentives for gaming in such systems. Future research may test this more directly. Future research may also attempt to separate more clearly effects of learning from performance measurement at the street-level (e.g., teachers, parents and students) from the effects of managers using the performance information to prioritize which performance goals to pursue (cf. Holm 2018). Also more detailed research on whether performance information work by making teachers update their prior beliefs about student skills – thereby enabling them more effectively to tailor their teaching to the needs of individual students would be interesting.

The positive results from the present studies make all such questions for future research all the more relevant to pursue.

References

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122(3): 1235–64.

Beuchert, L. V., M. K. Humlum, H. S. Nielsen and N. Smith (2018), "The Short-term Effects of School Consolidation on Student Achievement: Evidence of Disruption?" *Economics of Education Review* 65: 31-47.

Beuchert, L. V. and A. B. Nandrup (2018), "The Danish National Tests at a Glance." *Danish Journal of Economics* 1: 1-37.

Black, S. (1999), "Do better schools matter? Parental valuation of elementary education." *Quarterly Journal of Economics* 114: 577-599.

Black, P. and D, Wiliam (1998), "Assessmet and Classroom Learning." *Assessment in Education: Principles, Policy & Practice* 5 (1): 7-74.

Bond, T. G. and C. M. Fox (2007), *Applying the Rasch Model. Fundamental Measurement in the Human Sciences.* Second edition. Routledge.

Burgess, S., D. Wilson and J. Worth (2013), "A natural experiment in school accountability: The impact of school performance information on student progress." *Journal of Public Economics* 106: 57-67.

Chakrabarti, R. (2014), "Incentives and responses under *No Child Left Behind*: Credible threats and the role of competition." *Journal of Public Economics* 110: 124-146.

Danish Ministry for Children, Education and Gender Equality (2008), Fact Sheet: The Folkeskole: The "Folkeskole" is the Danish Municipal Primary and Lower Secondary School. Link.

Danish Ministry for Children, Education and Gender Equality (2011a), Kvalitetsrapporten som kommunalt styringsredskab (In English: The Quality Assessment as an Accountability Measure), Kontor for Kvalitetssikring og Kvalitetsudvikling, Skolestyrelsen. Link.

Danish Ministry for Children, Education and Gender Equality (2011b), De nationale test og kommunen - brug af testresultater i kommunens kvalitetsarbejde (In English: The municipality and the nationwide tests – applying test scores in the municipal quality assessment). Link.

Dee, T. D. and B. Jacob (2011), "The Impact of No Child Left Behind on Student Achievement." *Journal of Policy Analysis and Management* 30: 418-446.

Deming, D. J., S. Cohodes, J. Jennings and C. Jencks, (2016) "School Accountability, Postsecondary Attainment and Earnings." *Review of Economics and Statistics* 98(5): 848–862

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.

Figlio, D. and M. Lucas (2004), "What's in a grade? School report cards and the housing market." *American Economic Review* 94: 591-604.

Figlio, D. and C. E. Rouse (2006), "Do accountability and voucher threats improve low-performing schools?" *Journal of Public Economics* 90: 239-255.

Fitzpatrick, M. D., D. Grissmer and S. Hastedt (2011), "What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment." *Economics of Education Review* 30: 269-279.

Harris, D. N. (2011), "Value-Added Measures and the Future of Educational Accountability." *Science* 333: 826-827.

Hattie, J. (2009), *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Heinesen, E. (2010), "Estimating class-size effects using within-school variation in subject-specific classes." *Economic Journal* 120: 737-760.

Heinrich, C. J. and G. Marschke (2010), "Incentives and their dynamics in public sector performance management systems." *Journal of Policy Analysis and Management* 29 (1): 183-208.

Jacob, B. A. and S. D. Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3): 843-878.

Jacob, B. A. (2005), "Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89: 761-796.

Jussim, L., and K. D. Harber (2005), "Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies." *Personality and Social Psychology Review* 9: 131–155.

Krieg, J.M. (2011), "Which students are left behind? The racial impacts of the No Child Left Behind Act." *Economics of Education Review* 30 (4): 654-664.

Linn, R. L. (2001), "A century of standardized testing: Controversies and pendulum swings." *Educational Assessment* 7: 29–38.

Lipsey, M., K. Puzio, C. Yun, M. A. Hebert, K. Steinka-Fry, M. W. Cole, M. Roberts, K. S. Anthony, and M. D. Busick (2012), "Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms." NCSER, US Department of Education.

Neal, D. and D. W. Schanzenbach (2010), "Left Behind by Design: Proficiency Counts and Test-Based Accountability." *Review of Economics and Statistics* 92: 263-283.

Meadows, S., D. Herrick, A. Feiler and ALSPAC Study Team (2007), "Improvement in National Test Reading Scores at Key Stage 1: Grade Inflation or Better Achievement?" *British Educational Research Journal* 33: 47-59.

OECD (2004), OECD-rapport om grundskolen i Danmark. OECD.

Pøhler, L. and S. Sørensen (2010), Nationale test og anden evaluering af elevens læsning (In English: National Tests and other Evaluation of Reading Skills), *Dafolo Forlag*, 1st edition.

Rambøll (2013), Evaluering af de Nationale Test i Folkeskolen (In English: Evaluation of the National Tests in Elementary School), Rambøll Management Consulting.

Rambøll (2014), "Supplement til Evaluering af de Nationale Test i Folkeskolen." (In English: A Supplement to the Evaluation of the National Tests in Elementary School), Rambøll Management Consulting.

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

Reback, R. (2008), "Teaching to the rating: School accountability and the distribution of student achievement." *Journal of Public Economics* 92: 1394:1415.

Rockoff, J. E., D. O. Staiger, T. J. Kane, and E. S. Taylor (2012), "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review* 102(7): 3184–3213

Roediger, H. L., III, A. L. Putnam and M. A. Smith (2011), "Ten benefits of testing and their applications to educational practice." Ch. 1 in J. Mestre and B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education.* Oxford: Elsevier

Roediger, H. L., III, and J. D. Karpicke (2006), "The Power of Testing Memory: Basic Research and Implications for Educational Practice." *Perspectives on Psychological Science* 1: 181-210.

Rosenthal, R., and L. Jacobson (1968), *Pygmalion in the classroom*. New York: Holt, Rinehart, and Winston.

Rouse, C. E., J. Hannaway, D. Goldhaber and D. Figlio (2013), “Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure.” *American Economic Journal: Economic Policy* 5(2): 251-281.

Tenenbaum, H. R., and M. D. Ruck (2007), “Are teachers’ expectations different for racial minority than for European American students? A meta-analysis.” *Journal of Educational Psychology* 99: 253–273.

Appendix.

FIGURE A1

OVERVIEW OF TEST ACTIVITY OVER TIME, ALL TESTS

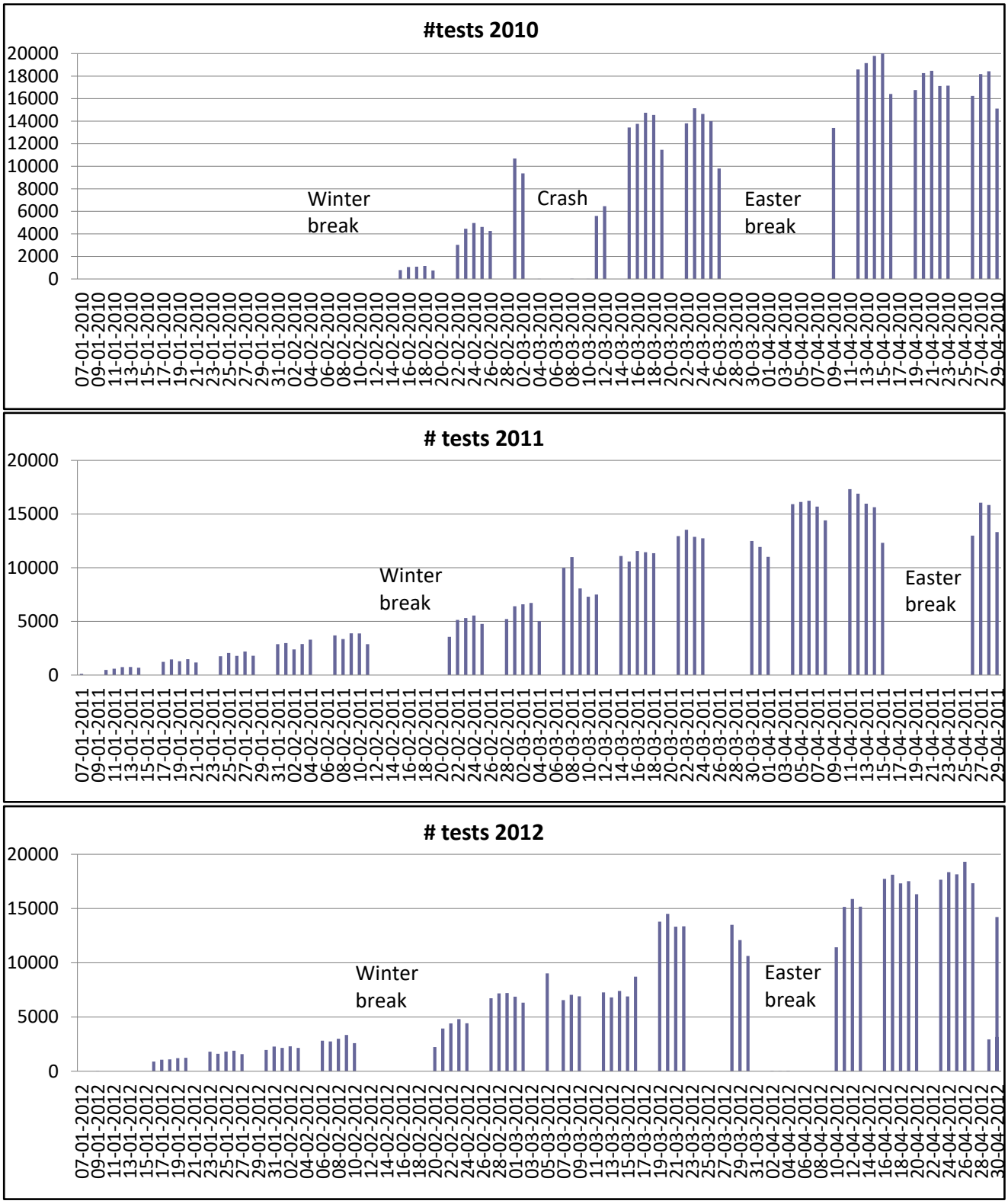
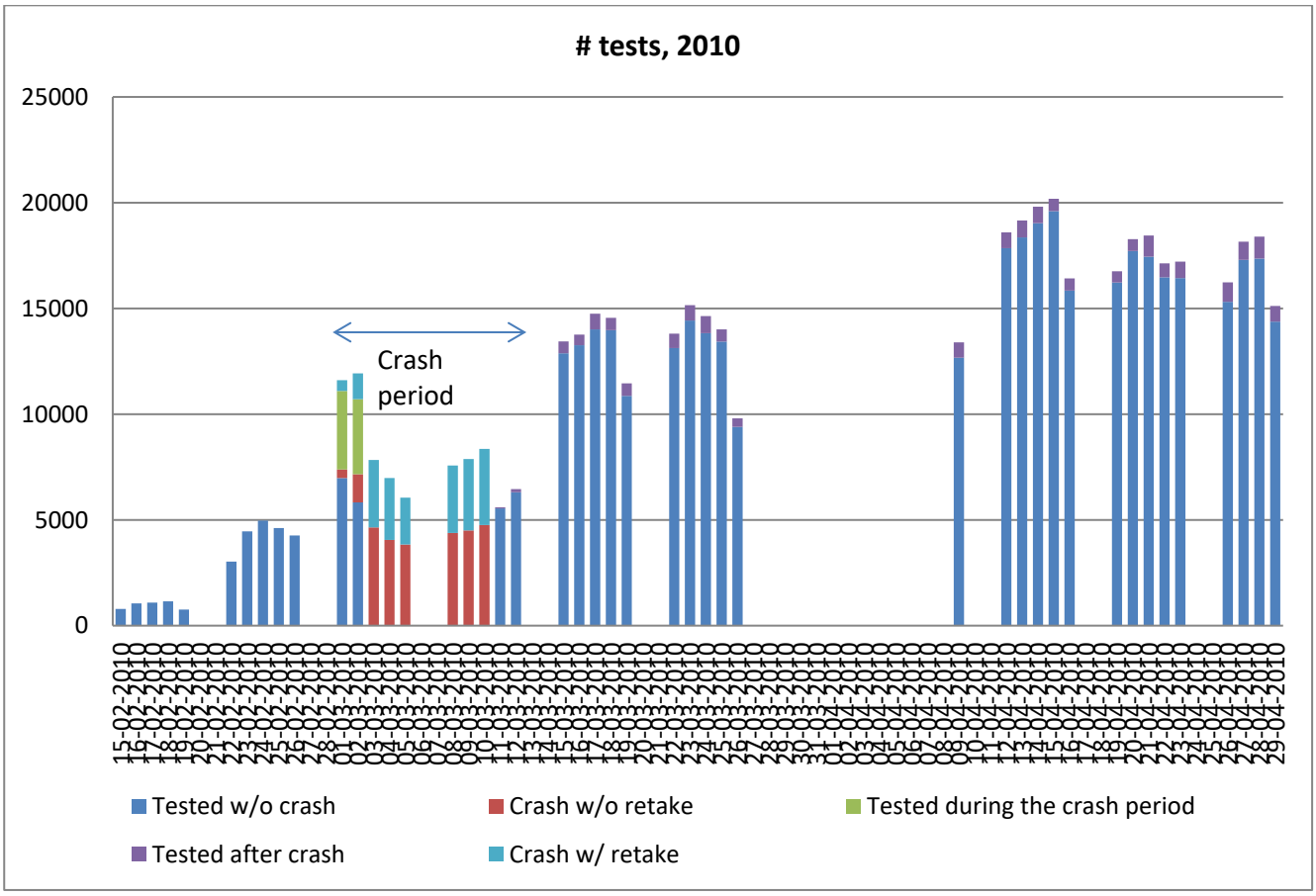


FIGURE A2

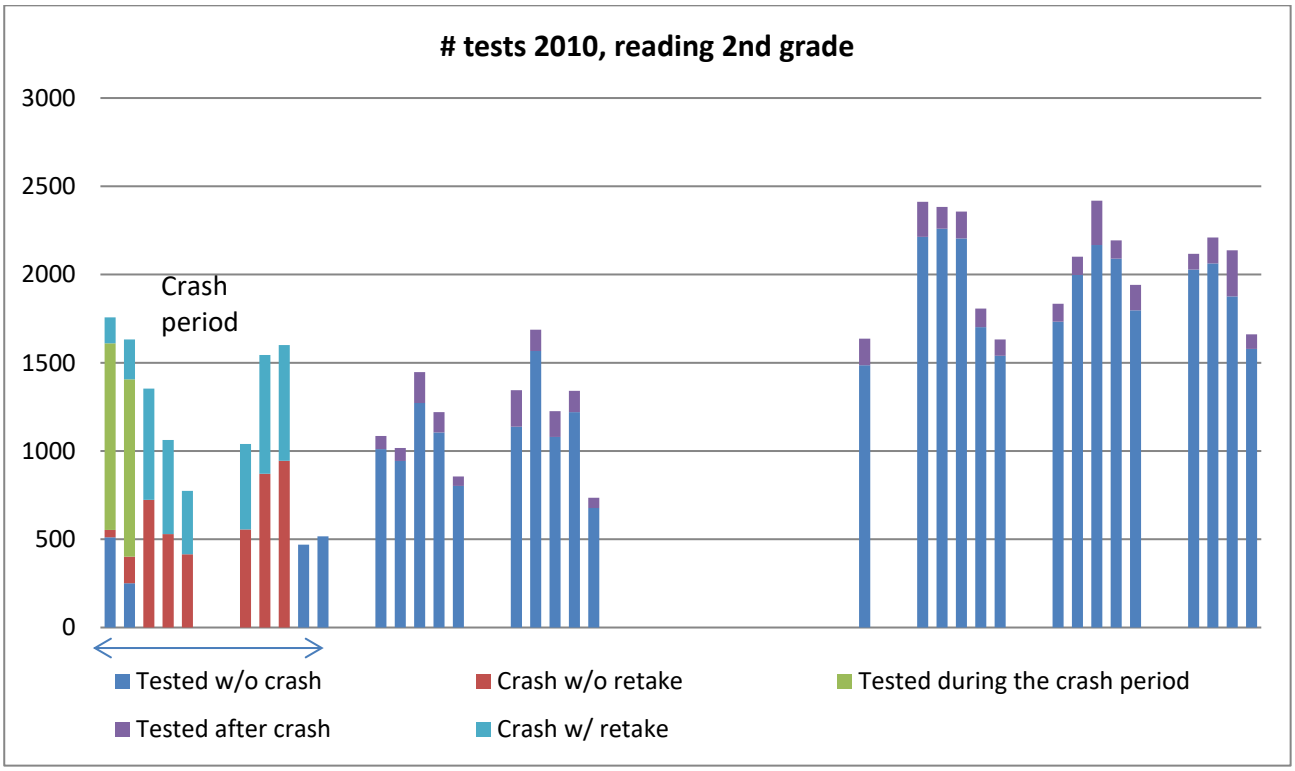
OVERVIEW OF TEST ACTIVITY OVER TIME, ALL TESTS

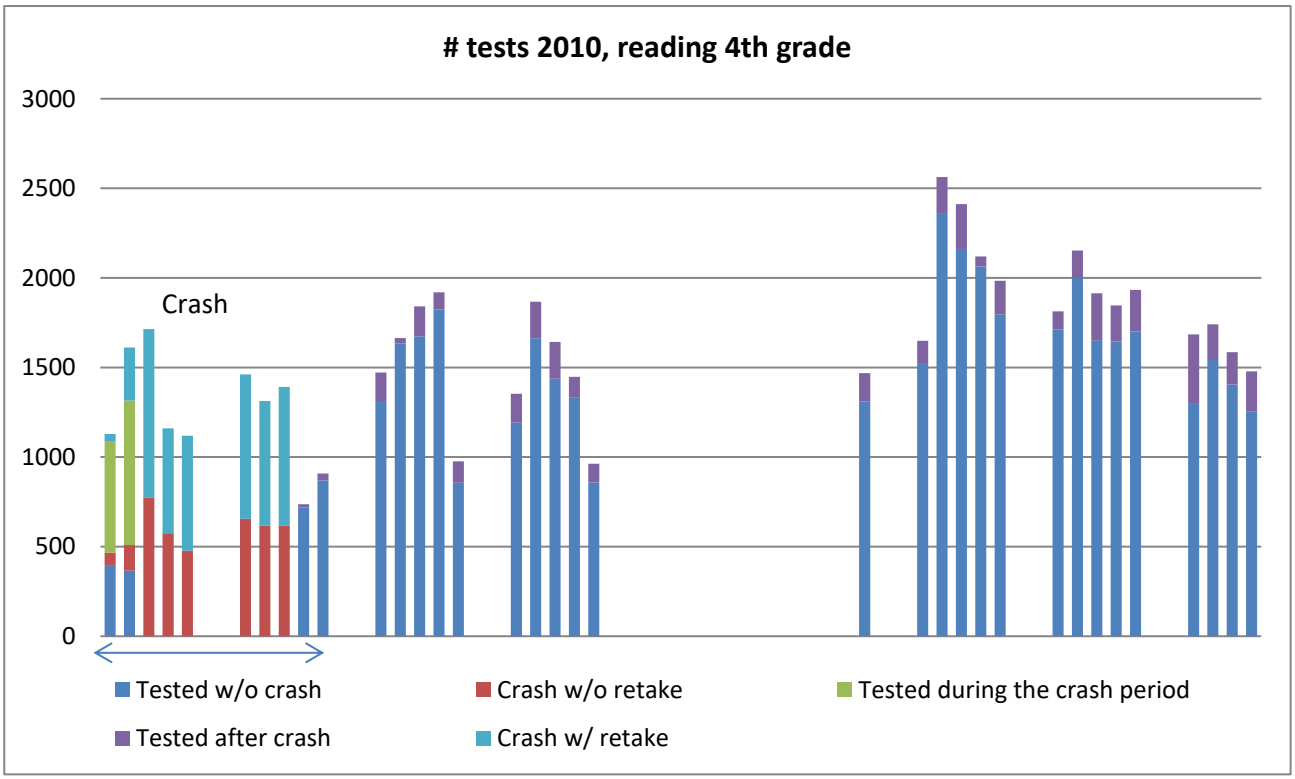


Note: March 1-2, 2010: The system was unstable and some students were recorded as booked during the crash and some were not, but most students completed the test. March 3-10, 2010: The system was completely shut down. March 11-12, 2010: The system was open but testing was voluntary; however, no records are available indicating which students were booked without completing the test.

FIGURE A3

OVERVIEW OF TEST ACTIVITY OVER TIME, READING 2ND AND 4TH GRADE

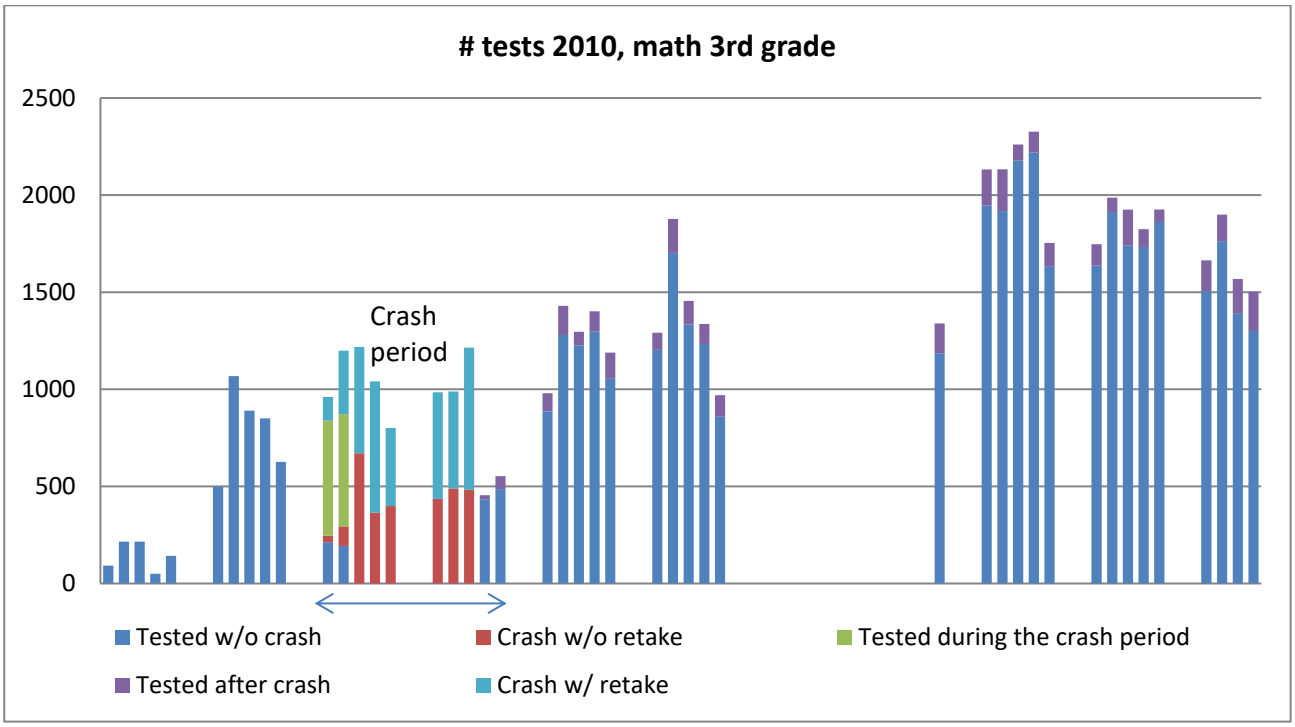


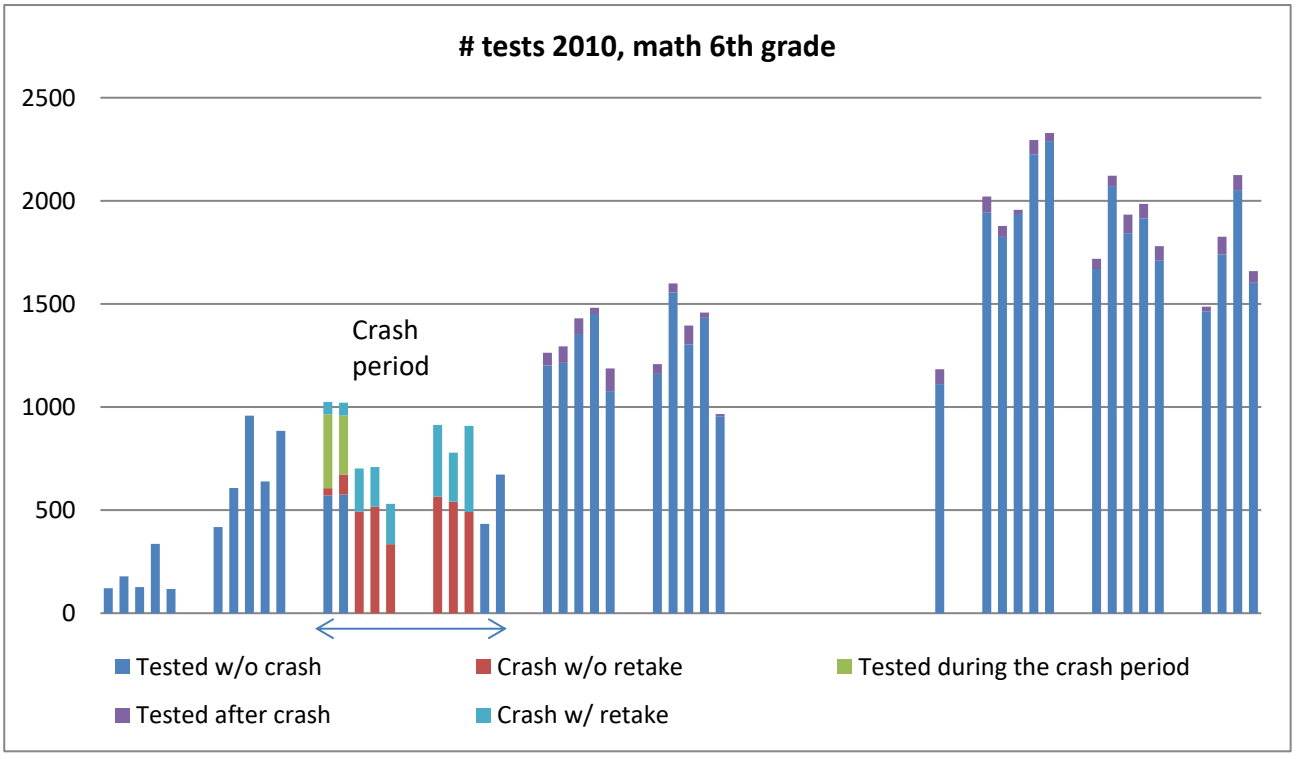


Note: See Figure A2.

FIGURE A4

OVERVIEW OF TEST ACTIVITY OVER TIME, MATH 3TH AND 6TH GRADE

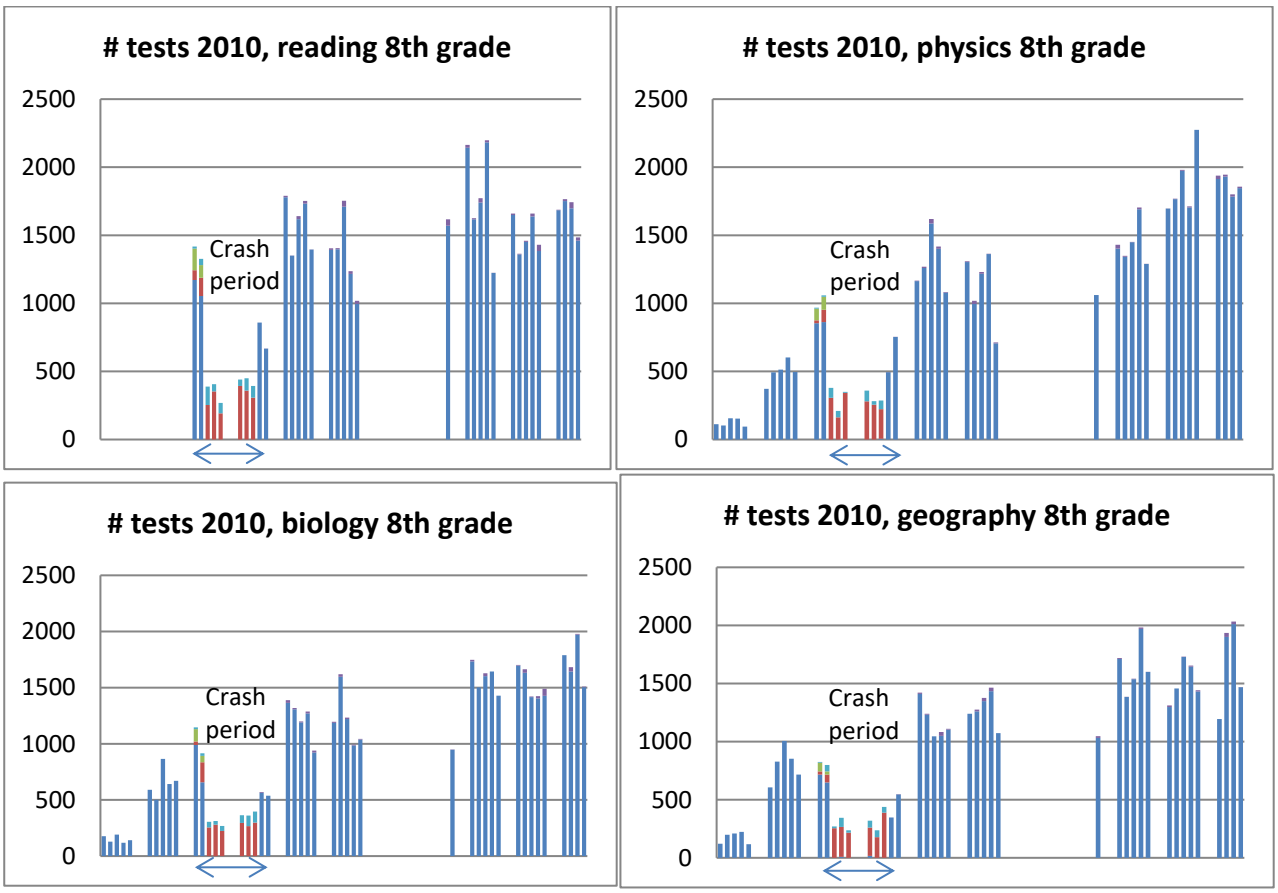




Note: See Figure A2.

FIGURE A5

OVERVIEW OF TEST ACTIVITY OVER TIME, 8TH GRADE



Note: See Figure A2.

TABLE A1

THE EFFECT OF TAKING A READING TEST IN 2ND, 4TH, AND 6TH GRADE ON READING PERFORMANCE TWO YEARS LATER

	(1) Total Score	(2) Total Score	(3) Total Score	(4) Language Comprehension	(5) Decoding	(6) Reading Comprehension
First Stage: Tested						
Crash	-0.502***	-0.502***	-0.503***	-0.503***	-0.503***	-0.503***
	(0.0211)	(0.0211)	(0.0210)	(0.0210)	(0.0210)	(0.0210)
Reduced Form: Test Score						
Crash	-0.0360**	-0.0358**	-0.0462***	-0.0436***	-0.0447***	-0.0317***
	(0.0153)	(0.0154)	(0.0107)	(0.0114)	(0.0104)	(0.0103)
Second Stage: Test Score						
Tested	0.0716**	0.0712**	0.0918***	0.0867***	0.0888***	0.0631***
	(0.0310)	(0.0311)	(0.0218)	(0.0231)	(0.0211)	(0.0209)
Second Stage (School FE): Test Score						
Tested	0.0597***	0.0560**	0.0517**	0.0220	0.0697***	0.0411*
	(0.0224)	(0.0223)	(0.0210)	(0.0215)	(0.0217)	(0.0215)
Number of Observations	136,887	136,887	136,887	136,887	136,887	136,887
Control Variables:						
Gender		X	X	X	X	X
+ grade		X	X	X	X	X
+ family background			X	X	X	X

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

TABLE A2

THE EFFECT OF TAKING A MATH TEST IN 3RD GRADE

ON MATH PERFORMANCE THREE YEARS LATER

	(1) Total Score	(2) Total Score	(3) Total Score	(4) Numbers and Algebra	(5) Geometry	(6) Applied Mathematics
First Stage: Tested						
Crash	-0.436***	-0.436***	-0.436***	-0.436***	-0.436***	-0.436***
	(0.0306)	(0.0306)	(0.0306)	(0.0306)	(0.0306)	(0.0306)
Reduced Form: Test Score						
Crash	-0.0343	-0.0343	-0.0303	-0.0303	-0.0219	-0.0270
	(0.0265)	(0.0266)	(0.0216)	(0.0225)	(0.0183)	(0.0209)
Second Stage: Test Score						
Tested	0.0794	0.0794	0.0701	0.0701	0.0508	0.0624
	(0.0610)	(0.0610)	(0.0497)	(0.0516)	(0.0422)	(0.0481)
Number of Observations	46,338	46,338	46,338	46,338	46,338	46,338
Control Variables:						
Gender		X	X	X	X	X
+ grade		X	X	X	X	X
+ family background			X	X	X	X

Note: Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1.

TABLE A3

HETEROGENEITY FOR TIMING OF TESTS:
THE EFFECT OF TAKING A READING TEST IN 2ND, 4TH, AND 6TH GRADE
ON READING PERFORMANCE TWO YEARS LATER

	(1) Benchmark	(2) Narrow Window around Crash	(3) Control for Late Subsequent Test
First Stage: Tested			
Crash	-0.503***	-0.503***	-0.502***
	(0.0210)	(0.0211)	(0.0210)
Relative to Overall First Stage		1.000	0.998
Reduced Form: Test Score			
Crash	-0.0462***	-0.0482***	-0.0405***
	(0.0106)	(0.0124)	(0.0106)
Second Stage: Test Score			
Tested	0.0918***	0.0958***	0.0807***
	(0.0218)	(0.0251)	(0.0216)
Number of Observations	136,887	58,012	136,887

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1. Narrow window is defined as +/- two weeks around the crash, i.e., before April 1, 2010. Control for late subsequent test is defined as tests taken April 1, 2010 or later.

TABLE A4

HETEROGENEITY FOR TIMING OF TESTS:
THE EFFECT OF TAKING A MATH TEST IN 3RD GRADE
ON MATH PERFORMANCE THREE YEARS LATER

	(1) Benchmark	(2) Narrow Window around Crash	(3) Control for Late Subsequent Test
First Stage: Tested			
Crash	-0.436***	-0.436***	-0.432***
	(0.0306)	(0.0306)	(0.0307)
Relative to Overall First Stage		1.000	0.991
Reduced Form: Test Score			
Crash	-0.0303	-0.0134	-0.0222
	(0.0216)	(0.0237)	(0.0213)
Second Stage: Test Score			
Tested	0.0701	0.0309	0.0514
	(0.0497)	(0.0546)	(0.0490)
Number of Observations	46,338	21,366	46,338

Note: All control variables are included. Standard errors are clustered at the school level, *** p<0.01, ** p<0.05, * p<0.1. Narrow window is defined as +/- two weeks around the crash i.e. before April 1, 2010. Control for late subsequent test is defined as tests taken April 1, 2010 or later.