Journal of Public Administration Research And Theory, 2019, 1–17
doi:10.1093/jopart/muz036
Article

Article

Learning from Performance Information

Simon Calmar Andersen*, Helena Skyt Nielsen†

*Department of Political Science, TrygFonden's Centre for Child Research, Aarhus University; †Department of Economics and Business Economics, TrygFonden's Centre for Child Research, Aarhus University

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 have begun to focus on situations in which performance management may facilitate internal organizational learning. In this study, we focus on one key component in performance management systems, namely generation of performance information. Based on a Bayesian learning model, we argue that generation of performance information at the individual level may create performance improvements because both users and frontline workers may learn where to prioritize their efforts. To test the isolated effect of this key component of any performance management system, we use as-good-as-random variation in exposure of students to testing because of a technical breakdown in an 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, and effects tend to persist after four years. We find larger effects for students with low socioeconomic status but also that schools with many students from this group are more reluctant to measure their performance. Implications and limitations in terms of increasing the level of student testing are discussed.

Introduction

Performance management is a multi-faceted phenomenon. Besides its core component of measuring individual units' performance, it consists of setting one or more organizational targets on one or more performance dimensions, evaluating the performance against these targets and using this evaluation to prioritize efforts (Andersen 2008; see also Boyne 2010; Moynihan 2008). Furthermore, performance management involves, to varying extents, the use of incentives to hold individuals and organizations accountable for their performance. Reviews of existing research on performance management show that performance information is often not used (Kroll 2013) and that the average effects of performance management systems are small and uncertain because of a 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. To move the research on performance management onwards in a constructive manner, we propose to study the individual components of

We are grateful for comments by Louise Beuchert, Helen Ladd, and Marianne Simonsen, as well as by participants at the APPAM conference and at the annual conference of TrygFonden's Centre for Child Research. We also want to thank the three reviewers and the editor for helpful and constructive comments that helped improve the manuscript. We appreciate Pernille Hansen's research assistance. Two reports (Rambøll 2013, 2014) funded by the Danish Ministry for Children, Education and Gender Equality present some of the results of this article. The Ministry, which provided inside information for the study design but did not influence the results, disclaims responsibility for the opinions expressed herein. The usual disclaimer applies. Address correspondence to the author at sca@ps.au.dk

performance management. Rather than embracing or discarding performance management altogether, understanding the effect of each performance management component would help us understand when and how performance management improves performance.

In this study, we focus on one key component, measuring individual users' outcome, which is often used as an organizational performance measure. We propose a Bayesian model of learning that explains why the measuring of individual performance itself may create performance improvements. In developing this theoretical argument, we build on work by Holm (2018), who argued theoretically and showed empirically how managers in a system with learning forums and repeated performance measures actually used performance information to prioritize low-performing areas and that these prioritizations resulted in improved performance over time. Relatedly, Jakobsen et al. (2018) argued that internal learning can be accomplished in systems that provide performance data to political principals if the high stakes are removed from the system (see also Andersen 2005). We argue that measuring outcomes for individual users may create learning in a Bayesian sense in which users and frontline workers use new performance information to update their prior beliefs about users' current state and use these new beliefs to prioritize their efforts.

Identifying the effects of performance management has proven difficult and often relies on nonexperimental data (Gerrish 2016; Kroll 2013). Isolating the effect of measuring individual users' performance, as we aim to do here, is even more difficult because there is often no variation in whether users are measured or not. Furthermore, the marginal effect of testing on the individual should be expected to be small relative to the baseline effect of all organizational efforts to improve performance. Estimating small effects with adequate statistical power requires very large sample sizes. To overcome these challenges, we exploit a technical breakdown during the introduction of an IT-based nationwide test system in Denmark involving 136,887 students. The technical breakdown affected an as-good-as-random subset of students. We use this variation generated by the breakdown to identify the students that were less likely to take the test and thereby less likely to learn from performance information.

Schools have often been used to test theories of management in general (O'Toole and Meier 2011) and performance management in particular (for a review, see Gerrish 2016; see also Snyder, Saultz, and Jacobsen 2017). As one of the largest groups of public organizations, the effect of performance management within schools is of interest in itself. Furthermore, the longstanding tradition of measuring student performance by test scores and exam grades make schools suitable for testing theories about performance management. However, research showing how the effect of public management within schools depends on the context of the schools (e.g., Meier et al. 2015a) makes us cautious in generalizing the empirical results from our study within education to performance management in general, and we encourage more studies in other policy fields. Furthermore, the performance management system we study is low-stakes systems relative to many Anglo-Saxon systems. If the effects of the individual components of performance management systems interact—especially if the effect of measuring individuals' outcome is different in high-stakes systems—results from the present study may not necessarily generalize. However, before studying any such contextual interaction effects, a first step in moving research on performance management forward will be to study the individual components of performance management.

To further explore the consequences of testing individual users' outcomes, we examine two additional research questions, namely (a) which groups of users benefit the most from being measured and (b) which organizations decide to measure individual users' outcomes. Results from these analyses show that students of low socioeconomic status (SES) tend to benefit more from being measured. This is consistent with the Bayesian learning model, if teachers' priors about low-SES students are more imprecise, as suggested by qualitative research (Harrits 2019; Harrits and Møller 2014). However, we also find that schools with more low-SES students are more reluctant to engage in measuring individual students' outcomes, with the unfortunate consequence that those who would benefit the most are least likely to receive the treatment.

In the next section, we present a theoretical framework explaining the effect of performance information within 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 and limitations of the results.

Theory and Existing Evidence

The performance management cycle involves a series of steps from deciding organizational goals (which in public organizations often involves multiple goals on various dimensions), selecting measures, measuring organizational performance (which often means measuring outcomes or satisfaction for individual users), evaluating the data, and deciding on how to prioritize efforts—before starting a new cycle of evaluation (Andersen 2008; Boyne 2010; Moynihan 2008). Empirical research has shown that on average, performance management systems do not improve performance. Gerrish (2016) conducted a meta-analysis of performance management studies and found a very weak relationship between this practice and outcomes. Forty-three percent of the studies in Gerrish's meta-analysis were conducted within education, which is more than four times as many as the second-most studied area. Yet, Gerrish's subgroup analysis of studies within education led to a similar conclusion of no or a weak relationship. Heinrich and Marschke (2010) reviewed the performance management literature within a principal–agent framework and pointed 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. The mixed results of existing performance management systems may be caused by any of the steps involved in the process as well as effects of interactions between them. One way forward for performance management research may therefore be to study the key components in isolation before studying their dynamic interplay. In this study, we focus on one key component, namely generating performance information.

The basic reasoning behind the effect of measuring performance follows a Bayesian learning model: Agents base their posterior beliefs on both their prior beliefs and any new information they receive. These updated beliefs should make agents act differently in their attempt to improve performance, and this reaction should improve outcomes for clients. However, we argue that at least two conditions are needed for performance information to have such effect. First, new information should differ from what agents already believe in order to create new learning. Second, agents should be motivated to react to updated, posterior beliefs. Even though we do not test these mechanisms and conditions, we believe it is helpful to make explicit the theoretical reasoning that guides the analysis and interpretation of the study.

Learning

The first condition for performance information to have an effect would be that agents learn from the information. If the new information does not differ from what agents already believe about their clients, we would not expect them to change their behavior, and we would therefore not expect to observe any effect of the new information.

Recent research in performance management suggests that agents interpret performance information relative to different reference points such as comparisons to historical performance, performance of similar organizations or politically set targets (e.g., Holm 2017). Interpreted within the Bayesian framework, these aspiration levels correspond to agents' prior beliefs. Meier, Favero, and Zhu (2015b) use a Bayesian learning model to develop hypotheses about how managers form their prior beliefs or expectations, and how they may react to gaps between their priors and new performance information. They argue that managers use the historical performance of the organization to form their baseline beliefs (reference points).[1] Holm (2018) suggests that public managers use information on different performance dimensions of their organization to strategically prioritize their efforts. In particular, he finds that managers prioritize performance dimensions with lowest performance (relative to a dimension's theoretical maximum). More related to our study, Rockoff et al. (2012) show in an experimental study how school principals use information about teacher performance to update their beliefs about the teachers and use these posterior beliefs to make personnel decisions, which results in more employee turnover and ultimately increased student performance. Additionally, they model how managers take the uncertainty involved in both their priors and new information into account when updating their beliefs.

We suggest that not just managers, but also frontline workers and service users (often acting as coproducers of the services) may in a similar way use performance information at the individual level to learn about where to prioritize their efforts. Empirical research within education indeed points to such learning effects among these different stakeholders.

As described in detail by Roediger, Putnam, and Smith (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, Putnam, and Smith 2011). Outside the lab, students may also learn from receiving teacher feedback based on the test results (Hattie 2009; Hattie and Timperley 2007). If standardized testing is used for formative assessment (and not just a summative assessment), where student performance is compared with a reference level (which may function as the prior beliefs in the Bayesian learning model; Meier, Favero, and Zhu 2015b), and test results will be used for taking actions to fill the gap (i.e., the difference between prior beliefs and the

[1] In O'Toole and Meier's (1999 classical model of public management, the relationship between past performance and future performance is explicitly modeled by an autoregressive component, even though they do not specify how this variable level of stability or inertia may depend on how managers learn from information of past performance relative to any prior beliefs they may have.

new information generated by the test), 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. In an experimental study Dizon-Ross (2019) shows that providing parents with information on their children's performance in school causes the parents to update their beliefs and adjust their investments in the education of their children. Asking teachers to give students' test results to parents may also 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 (Andersen and Nielsen 2016; Jakobsen and Andersen 2013).

Finally, teachers themselves may learn from the test results. Randomized controlled trials testing interventions that use test results to target teaching to students' skill levels have shown very positive results (Banerjee et al. 2007). Teachers' prior beliefs about students may be imprecise or biased. Student testing may help teachers update their beliefs based on their priors and the new information, and use their updated beliefs to tailor their instructions to the individual students. In a new study, Bergbauer, Hanushek, and Woessmann (2018) find that school systems that introduce standardized, comparable test regimes experience improvements in student performance, whereas teachers' own student testing does not have any discernible effects.

Motivation

The motivation condition relates to studies of high- and low-stakes performance management systems. In high-accountability systems, principals use incentives such as publicizing performance information and rewarding high-performance scores to increase agents' motivation to react to the performance information (Muller 2018). Empirical research shows that high-accountability systems have 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 United States, indicate that they 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., Chakrabarti 2014; Deming et al. 2016; Figlio and Rouse 2006; Krieg 2011; Neal and Schanzenbach 2010; Reback 2008), whereas 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 more promising effects of accountability pressure on student achievement. Rouse et al. (2013) provide evidence that schools given the lowest grading significantly change their instructional policies and practices and that these responses explain a substantial part of subsequent test score gains. Based on a study of the impact of the National Assembly abolishing national testing in Wales, Burgess, Wilson, and Worth (2013) strongly advocate test-based accountability. Carnoy and Loeb (2002) also find that students improve more in math in high-accountability US states (see also Hanushek and Raymond 2005). Perhaps most relevant to our study, Dee and Jacob (2011) use a comparative, interrupted time series approach based on all US states and find significant effects of accountability on a low-stakes math test 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 fourth graders nevertheless fall below national proficiency standards, and there is still no robust effect on reading proficiency.

These mixed results indicate that high-stakes systems motivate agents to improve performance, but some of the improvements seem to be driven by strategic responses, which have led researchers to consider whether agents would be motivated to react to performance information in the absence of high-stakes incentives. This interpretation is in line with Boyne's (2010) proposition that the expected positive relationship between setting targets and performance is moderated by managerial gaming.[2] He further suggests that some level of involvement of employees in setting targets may improve motivation and reduce gaming—but also that handing too much control over to employees may result in another form of gaming, namely setting too easy targets.

In line with this reasoning, Jakobsen et al. (2018) argue that involving professionals in the interpretation of performance goals will increase motivation and,

[2] This also relates to Hood's (2011) analyses of how blame avoidance behavior increases as a function of how much the management of organizations (such as the use of performance management) promotes the assignment of responsibility of perceived avoidable harm.

ultimately, performance. Other research also suggests that public managers and employees are motivated to improve their own (organizational) performance, even without strong economic incentives or sanctions (e.g., Andersen, Heinesen, and Pedersen 2014; Meier, Favero, and Zhu 2015b). The motivation of service users (and often service coproducers) is more straightforward because they may have a self-interest in improving their own or close relatives' (such as their own children's) performance. Low-stakes tests (such as the Northwest Evaluation Association's so-called MAP tests) have become widespread in the United States, and this development may point to a belief that low-stakes tests will improve learning in schools.

For our study, it suffices to say that also in (relatively) low-stakes accountability systems may frontline workers as well as service users and coproducers possess enough motivation for performance improvement to react to what they learn from performance information. In the discussion, we return to the question whether the effect of performance information may depend on the accountability system in which it is generated.

Subgroup and Contextual Effects

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

Based on Bourdieu's (1984) connection between social differences and the use of social categorizations, Harrits (2019) argues and find in a qualitative study that that street-level bureaucrats with middle-class background tend to rely more on stereotypes when assessing clients of low SES than when they assess clients with their own social background (see also Harrits and Møller 2014). Such use of stereotypes suggests that teachers' prior beliefs about low-SES students would be more imprecise. This is also what Andersen, Guul, and Humlum (2018b) find in a recent study that compares teachers' perceptions of students' skills to performance information from standardized, norm-based test results calculated by computer algorithms.

Following the Bayesian model of learning, teachers' posterior beliefs would be more affected by performance information the more the new information diverges from their priors. So if frontline workers' prior beliefs about low-SES clients are more imprecise than for middle- and high-SES students, the Bayesian model suggests that effects of providing frontline workers with new information would be larger for these low-SES clients.

Second, we examine which schools comply with the performance measurement policy. Building on self-affirmation theory, Petersen, Laumann, and Jakobsen (2019) argue that when frontline workers are confronted with performance information that indicates that they have performed poorly, they are likely to engage in defensive biases to protect their self-image integrity. Such defensive biases can be a problem because they can prevent the frontline workers from learning from potentially important information (see also Sherman and Cohen 2006). A survey experiment among high-school teachers confirms Petersen, Laumann, and Jakobsen's (2019) expectation that the teachers are more reluctant to take responsibility for the results and more critical of the performance indicators when the test scores of their school are low.

This reasoning suggests that the effect of performance measurement may depend on the context in the sense that organizations expecting high performance would be more inclined to measure performance. More specifically in our educational context: Because schools with high shares of students with low SES on average have lower test scores, we expect schools with high proportions of students with low SES to be more reluctant to comply with the test regime.

We do not want to overstate the claim that testing increases learning. The specific effects of such a performance management system may depend on 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 eighth grade (approximately age 15). However, based on poor Programme for International Student Assessment results in 2000 and 2003, a subsequent OECD (2004) report, 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 and 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.[3] The initial accountability system comprised nationwide testing of students 10 times from second to eighth grades, annually updated individual plans for students, compulsory ninth-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 (Hess, Holmsgaard, and Jaokobsen 2009).

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. Each school is endowed with an annual socioeconomic index based on sex, ethnicity, and parental SES, enabling schools to compare test results with the national average at schools with a similar socioeconomic index. Schools are informed about the gap between the national average and their test score for each of the 10 tests and given an indication of whether the gap is statistically significant or not. School principals and parents are also supposed to be informed by the individual students' test results, which may generate some accountability pressure on the teachers. However, no explicit proficiency standard had been set for local authorities, schools, subgroups, or individuals at the time of our data generation.[4] Therefore, relative to systems with strong economic incentives or threats of firing teachers and closing schools because of low performance, the Danish system can be characterized as low-stakes

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 any significant deviance from other schools with similar socioeconomic indices but also measures taken to deal with unsatisfactory results. Poor academic performance does not result in automatic repercussions. This is not to say that there are no stakes in the Danish accountability system. For instance, 15% of students were affected by school consolidations during the 2010–11 and 2011–12 school years (Beuchert et al. 2018). During the political process of deciding such school consolidations, school performance may be considered. However, there is no direct relationship between test results and rewards or sanctions as found in many high-stakes accountability systems.

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 ninth-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 of the quality assessment, the future objectives of the school, and the local authorities as a whole. However, the average test score of schools is confidential. It is illegal to make results of individual students, classrooms, or schools publicly available in any way (Danish Ministry for Children, Education and Gender Equality 2011). This law prevents the media from publishing test results and league tables, and parents from selecting schools based on test results.

In sum, there are some elements of accountability tied to the student testing system in Denmark, but because it is illegal to publish the results of individual students, classrooms, schools, or municipalities and because there are no direct rewards or sanctions tied to the test results, it is a low-stakes accountability system—compared with systems in which measures of performance are tied to publishing and rewarding (cf., Muller 2018).

The Testing System

Nationwide testing was initially legislated in Denmark when the nationwide school accountability system was introduced in 2006. However, because of technical and other challenges, testing was postponed until the school year 2009–10. Beginning that year, 10 compulsory standardized national tests were introduced in the Danish public schools from second to eighth grades. Students were tested on various subjects throughout their school career, but the emphasis is on reading and math.[5] Both before and after the introduction of the national testing system, teachers, schools, and municipalities have used student tests. We test the effect of this new, additional performance information generated by the national testing program.

The compulsory tests take place from January to April at the end of grades 2–8, 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

[3] Detailed in the Public School Law (Law no. 313, April 19, 2006, and Law no. 572, June 9, 2006) and described by Hess et al. (2009). Within the framework of the national law, public schools in Denmark are governed by and accountable to local authorities.

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

[5] See Beuchert and Nandrup (2018) for a detailed description of Danish national tests.

Table 1. Compulsory National Tests

	Grade						
Subject	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth
Reading	X		X		X		X
Math		X			X		
English						X	
Geography							X
Physics/chemistry							X
Biology							X

pool of questions, but 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.

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 of 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 Supplementary 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, whereas the math test centers on numbers and algebra, geometry, and applied mathematics. 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 grading.

The tests are adaptive, which means they adapt to the child's abilities during the test, challenging children according to their skill level. The test begins with a moderately difficult question. If the question is answered correctly, the student receives a more difficult question, but if the question is answered incorrectly, the student receives an easier question, and so on. When the test results for all three domains reach a sufficiently statistical certainty, the test ends. The adaptive system means that teachers do not know the specific questions each student receives or the answers to these questions. Teachers only know the type of questions used in the test. They can use their knowledge for teaching-to-the-test but cannot game the system by providing students with correct answers to specific questions in advance.

The test scales measure the absolute ability within each domain. Approximately 183,000 individual test results, or 15,000–21,000 results for each of the 10 national tests, were used to score the tests when they were introduced in 2010 (Beuchert and Nandrup 2018). These results set the norm for the scale. The test has subsequently been revised 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 the 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 in 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

[6] Data on the timing of tests show 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.

[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."

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 follows:

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

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, whereas 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 period 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 to 2, 2010, and from March 11 to 12, 2010, because accurate information about who the IT problems affected is not available.

We exploit the fact that the crash—and the sudden exemption from taking the otherwise compulsory test—were unexpected for the students and the teachers. Teachers and students exposed to the crash were no less prepared for the test compared with the rest of the population. Some teachers rebooked some of the students who 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 instrumental variable (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 to estimate the parameters of interest. The IV analysis identifies the effect of taking the test for the "compliers," that is, the 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.

Table 2. Performance of the National Test IT System in 2010

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

Threats to Validity

Two channels for potential nonrandom selection of student testing could threaten the validity of the instrument: Selection in booking of test sessions and selection in exposure to the crash.

Nonrandom 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 Supplementary 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 second and fourth grades, and in Figure A4 for the math test in third grade. However, Figure A5 clearly shows that the information for eighth 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 (second–sixth grades) and for the early math test (third 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 we therefore 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 between March 3 and 10, 2010.

Nonrandom 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 robustness 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 the annual progress, the test window amounts to almost one-third of the school year, which could bias the results (see, e.g., Fitzpatrick, Grissmer, and Hastedt 2011). As an additional robustness 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 between 1996 and 2002.[8] This data set includes all students in second–sixth grades for the 2009–10 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. Because of 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 a mean of zero and an SD 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.[9] Table 3 shows that in these two samples, more than 75% of the students were unaffected by the crash and took the test as required. Five to seven percent 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.[10] 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.[11] These noncompliers are the reason behind our IV design.

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%

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 nor 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. To study heterogeneity across SES we split the sample by education and income. High education is defined as having at least one parent who completed a college education (45%) and low education as no parents with a college education (55%), whereas high versus low income is defined as income above or below DKK 220,000 (70% above and 30% below, respectively). Correlation matrices between key variables that are used to split the sample and test scores

[8] The data sets we use belong to the Danish Ministry of Children and Education and Statistics Denmark. Because data contain sensitive information on residents, they cannot be made publicly available. Researchers can, however, apply to Statistics Denmark and the Ministry of Children and Education for access.

[9] 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 with 91.0% of those not exposed to a crash.

[10] A few of these individuals could be returning from abroad or from private schools (<1%).

[11] For instance, the teachers' union represents a critical voice in this debate (see folkeskolen.dk).

Table 4. The Effect of Taking a Reading Test in Second, Fourth, and Sixth Grades on Reading Performance Two Years Later

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Males	Females	Low Education	High Education	Low Income	High Income	Non-Western Immigrants
First stage: tested								
Crash	−0.503	−0.506	−0.500	−0.502	−0.504	−0.516	−0.498	−0.505
	(0.0210)	(0.0214)	(0.0213)	(0.00199)	(0.00220)	(0.00273)	(0.00178)	(0.00610)
Relative to overall first stage		1.006	0.994	0.998	1.002	1.026	0.990	1.004
Reduced form: test score								
Crash	−0.0462	−0.0546	−0.0380	−0.0424	−0.0503	−0.0667	−0.0375	−0.0750
	(0.0106)	(0.0135)	(0.0120)	(0.00921)	(0.00916)	(0.0131)	(0.00758)	(0.0296)
	[<.0001]	[.0001]	[.0015]	[<.0001]	[<.0001]	[<.0001]	[<.0001]	[.0113]
Second stage: test score								
Tested	0.0918	0.108	0.0759	0.0845	0.0997	0.129	0.0753	0.149
	(0.0218)	(0.0273)	(0.0245)	(0.0265)	(0.0253)	(0.0331)	(0.0229)	(0.0713)
	[<.0001]	[.0001]	[.0019]	[.0014]	[.0001]	[.0001]	[.0010]	[.0392]
Number of observations	136,887	68,746	68,141	75,345	61,542	39,593	94,642	7,875

All control variables are included. Number of observations are different compared with earlier since not all schools have ninth grade. Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and *p*-values in brackets.

are shown in Tables A1 and A2 in Supplementary Appendix.

Summary statistics are given in Table A3, which shows means for background variables by crash status (unaffected or exposed) for the estimation samples. Differences across crash status are small in magnitude, and only a few are statistically significant.[12] 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 4, 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 an SD. The second-stage results show that taking a reading test increases the test score as measured two years later by about 9% of an SD.

Table A4 in Supplemenary 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.[13]

Table 5, 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 therefore slightly higher than for reading. The point estimate in the second-stage results is 7%, which is similar to the point estimate for reading, although insignificant because of the smaller sample size. Table A5 in Supplemenary Appendix A shows that the results are almost unaffected when controls are added and that they do not vary much across profile areas.

The reason for the difference between the effects on math and reading may be that math was measured

[12] The pattern is similar when we zoom in on students booked before April 1. In the empirical analyses, we present robustness analyses in which we study a narrow test window around the crash.

[13] Table A8 shows results for reading that include school fixed effects to examine how much of the total effect is due to within-school variation. The estimates are roughly halved when school fixed effects are added, but they are still statistically significant. This suggests that the effect is partly driven by school variation (such as school management and test culture), partly by responses particular to the student's test experience or the subsequent student–teacher–parent interaction. Adding school fixed effects to the math estimations does not make much sense because only one cohort is included in our study compared with three cohorts for reading.

Table 5. The Effect of Taking a Math Test in Third Grade on Math Performance Three Years Later

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	All	Males	Females	Low Education	High Education	Low Income	High Income	Non-Western Immigrants
First stage: tested								
Crash	−0.436	−0.441	−0.431	−0.460	−0.408	−0.457	−0.428	−0.512
	(0.0306)	(0.0319)	(0.0323)	(0.00328)	(0.00347)	(0.00464)	(0.00282)	(0.00998)
Relative to overall first stage		1.011	0.989	1.055	0.936	1.055	0.936	1.174
Reduced form: test score								
Crash	−0.0303	−0.0273	−0.0322	−0.0472	−0.00941	−0.0571	−0.0232	−0.0371
	(0.0216)	(0.0265)	(0.0253)	(0.0163)	(0.0188)	(0.0240)	(0.0145)	(0.0525)
	[.1607]	[.3029]	[.2031]	[.0038]	[.6167]	[.0174]	[.1096]	[.4798]
Second stage: test score								
Tested	0.0701	0.0624	0.0755	0.103	0.0233	0.126	0.0548	0.0717
	(0.0497)	(0.0586)	(0.0597)	(0.0484)	(0.0703)	(0.0641)	(0.0554)	(0.107)
	[.1584]	[.2869]	[.2048]	[.0333]	[.7403]	[.0493]	[.3225]	[.5028]
Number of observations	46,338	23,241	23,097	24,754	21,584	12,228	33,254	2,617

Note: All control variables are included. Number of observations are different compared with earlier since not all schools have ninth grade. Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and *p*-values in brackets.

after three years and reading after two years, and the effects may have faded out. Table 6 supports this interpretation. The table shows the effect on reading four years after the initial tests. Effect sizes decreased from .09 after two years to about .07 after four years, but they are still statistically significant. This is close to the math results, which are estimated to be around .07 after three years. The math results are estimated more imprecisely (partly) due to smaller sample size. These analyses also suggest that short-term effects may be even larger.[14]

Heterogeneity across Groups of Students

The results above suggest that test exposure is positively related to subsequent test performance. Tables 4 and 5 (columns 2–8) show the results divided by gender, parental education and income, 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 higher for low-income parents (<220,000 DKK) than for high-income parents. The point estimate of being tested in math is high and significant for students with low parental education and income but close to zero for students with high education and income. These results are evidence against the concerns that weaker students are harmed by testing (Deming et al. 2016) but aligns with our expectation that effects would be larger for this group of students because teachers may have had more imprecise perceptions of their skills prior to the testing.

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 ninth-grade exams in the 2008–09 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 7 shows the results. The results are consistent with Petersen, Laumann, and Jakobsen's (2019 finding that teachers with low test scores are less likely to take responsibility for test results.

In Tables 8 and 9, we examine whether the effects of testing differ for schools with high shares of

[14] We also examined whether results differ by grade level, but found no systematic differences. Similarly, we examined whether testing in reading had spill-over effects on math and vice versa but did not find much evidence of this.

Table 6. The Effect of Taking a Reading Test on Reading Performance Four Years Later

	(1)
First stage: tested	
Crash	−0.487
	(0.0217)
Reduced form: test score	
Crash	−0.0346
	(0.0134)
	[.0098]
Second stage: test score	
Tested	0.0711
	(0.0276)
	[.0100]
Number of observations	85,676

All control variables are included. Number of observations is smaller compared with Table 4 because only individuals in second or fourth grade in 2010 can be included in a four-year follow-up. Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and *p*-values in brackets.

disadvantaged students. First, we analyze the compliers to characterize the population who responds to the instrument and provides 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), whereas 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, whereas it is literally zero for students attending schools in the highest decile. This may reflect that these schools already have sufficiently 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, in which 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 being important for the impact of being tested.

Robustness Checks

Tables A6 and A7 in Supplemenary Appendix present robustness checks with respect to the timing of the tests. Column 1 reproduces the main results from Tables 4 and 5, whereas 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 not driven by the timing of subsequent tests.

Concluding Discussion

Evaluations of performance management systems tend to find on average no or small effects on organizational performance. However, this average covers a substantial variation across studies with some studies pointing to positive effects, whereas others find outright negative and unintended effects. In this study, we propose that one way forward is to study the different components of performance management separately to enhance our theoretical understanding of how each component may contribute to performance improvements. We find that one core component of performance management systems—generating and conveying performance information to frontline workers—by itself is enough to enhance performance. Our study suggests beneficial effects of testing the students of around 0.09 of an SD on reading test scores two years later, and these effects are only slightly reduced after four years. Math test scores are measured after three years and show similar effect size.

The effect sizes are about three times higher than what Gerrish (2016) found in his meta-analysis of performance management studies (overall effect size of 0.03 and 0.02 for studies within education). This difference, however, may be due to the use of correlational evidence in many existing studies of performance management. The magnitude we find is similar to

Table 7. Regression of Indicators of Test Behavior on School Characteristics

	Sign up for Compulsory Test in 2010			
	(Sample: All)			
	(1)	(2)	(3)	(4)
Reading second, fourth, and sixth grades				
Lowest decile exit exams		−0.0370	0.00376	0.00908
		(0.0127)	(0.0149)	(0.0147)
		[.0036]	[.8263]	[.5368]
Highest decile exit exams		0.0208	0.00965	0.0118
		(0.00766)	(0.00852)	(0.00852)
		[.0066]	[.2569]	[.1661]
Missing exit exams		−0.0138	−0.00691	−0.00626
		(0.00661)	(0.00647)	(0.00641)
		[.0368]	[.2848]	[.3288]
Lowest quarter SES	−0.0671		−0.0668	−0.0540
	(0.00736)		(0.00742)	(0.00721)
	[<.0001]		[<.0001]	[<.0001]
Highest quarter SES	−0.00732		−0.00912	−0.0103
	(0.00637)		(0.00688)	(0.00688)
	[.2506]		[.1845]	[.1344]
More than 15% immigrants	−0.0323		−0.0332	−0.0188
	(0.0116)		(0.0121)	(0.0118)
	[.0068]		[.0082]	[.1111]
Constant term	0.927	0.907	0.928	0.924
	(0.00381)	(0.00353)	(0.00416)	(0.00752)
	[<.0001]	[<.0001]	[<.0001]	[<.0001]
Additional controls				X
Mean of dep. variable	0.9043			
Number of observations	151,375			
Math third grade				
Lowest decile exit exams		−0.0252	0.0131	0.0191
		(0.0153)	(0.0208)	(0.0205)
		[.0995]	[.5288]	[.3515]
Highest decile exit exams		0.0187	0.0233	0.0313
		(0.0145)	(0.0159)	(0.0159)
		[.1972]	[.1428]	[.0490]
Missing exit exams		−0.150	−0.146	−0.134
		(0.118)	(0.114)	(0.100)
		[.2037]	[.2003]	[.1802]
Lowest quarter SES	−0.0756		−0.0603	−0.0430
	(0.0397)		(0.0241)	(0.0135)
	[.0569]		[.0123]	[.0014]
Highest quarter SES	−0.0239		−0.0357	−0.0325
	(0.00944)		(0.0126)	(0.0119)
	[.0113]		[.0046]	[.0063]
More than 15% immigrants	−0.0332		−0.0518	−0.0383
	(0.0134)		(0.0193)	(0.0169)
	[.0132]		[.0073]	[.0234]
Constant term	0.922	0.926	0.955	0.946
	(0.0220)	(0.00459)	(0.00972)	(0.0140)
	[<.0001]	[<.0001]	[<.0001]	[<.0001]
Additional controls				X
Mean of dep. variable	0.8932			
Number of observations	51,880			

Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and p-values in brackets. SES, socioeconomic status.

Table 8. Heterogeneity by School Characteristics: The Effect of Taking a Reading Test in Second, Fourth, and Sixth Grades 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-Western Immigrants
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)
	[.0115]	[.8982]	[.0001]	[.0003]	[<.0001]
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)
	[.1156]	[.9273]	[.0185]	[.0283]	[.0096]
Number of observations	5,222	8,024	33,840	35,160	12,779

All control variables are included. Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and *p*-values in brackets. SES, socioeconomic status.

Table 9. Heterogeneity by School Characteristics: The Effect of Taking a Math Test in Third 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-Western Immigrants
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)
	[.9134]	[.2526]	[.1846]	[.9663]	[.0092]
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)
	[.9501]	[.4913]	[.4205]	[.9983]	[.1038]
Number of observations	1,833	2,726	10,596	13,915	5,030

All control variables are included. Outcome variables are standardized to have mean of zero and standard deviation of one. Standard errors clustered at the school level are reported in parentheses and *p*-values in brackets. SES, socioeconomic status.

what has been found in studies using stronger identification strategies. According to Figlio and Ladd (2015) studies exploiting discontinuities in accountability systems typically find effect sizes of 0.05–0.10 SD. Burgess et al. (2013) find an effect of 0.07 SD using a difference-in-differences model for United Kingdom, whereas Deming et al. (2016) exploit grade-cohort variation in accountability pressure in Texas public schools and find an average effect of 0.05 SD on tenth-grade score.

On average, student progress in reading for an entire school year is 32%–40% of an SD during the relevant grades (Lipsey et al. 2012).[15] Taking the test thereby increases what a student would have learned over a two-year period by an estimated 12.5%.[16] On

[15] 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 an SD in reading and 52%, 56%, and 41% in math.

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

average, student progression in math over the three school years (grades 3–6) is 150% of an SD (Lipsey et al. 2012). Taking the test therefore increases what a student would have learned in over a three-year period by an estimated 5%. The effect sizes are also comparable to the effect of reducing class sizes by three to four students (Heinesen 2010) or having a coteacher in the classroom most of a school year (Andersen et al. 2018a), which are much more costly policies.

Implications for Theory

We interpret the results within a Bayesian learning model. Our study does not test the individual components of the model, that is, the prior beliefs and the updating of beliefs based on new information, but results are consistent with this model. In particular, we find—in accordance with other research showing that teachers' beliefs about low-SES students tend to be more imprecise—that low-SES students tend to benefit more from being tested. We believe that our study, combined with theoretical developments within public administration (Meier, Favero, and Zhu 2015b), and empirical studies of school principals (Rockoff et al. 2012) and teachers (Andersen, Guul, and Humlum 2018b), contributes to demonstrating the relevance of the Bayesian learning model, even though our study itself does not prove this model. The Bayesian learning model suggests that performance information should not be seen or used as definitive answers on organizational or individual performance, let alone providing definitive answers on how managers or frontline workers should react to performance information. New information is interpreted in light of prior information. Yet, new information may help stakeholders update their prior beliefs, and especially when they diverge from the new information, there are reasons to reconsider whether efforts should be redirected to other performance dimensions or other service users. Consistent with Holm's (2018 finding that managers tend to prioritize performance dimensions that diverge most from expectations, our results suggest that performance information may be especially valuable for low-SES students, about whom teachers may have imprecise prior beliefs, and who generally have lower performance in school than high-SES students.

For that reason, it is unfortunate—but consistent with prior work on teachers' acceptance of performance information (Petersen, Laumann, and Jakobsen 2019)—that we find that schools with high shares of low-SES students (and thus lower performance scores)—are more reluctant to adopt the performance measurement system. These schools would, all else being equal, benefit the most from using the system.

Whereas Bayesian models of learning focus on the interplay between prior beliefs and new information they have focused less on which conditions are needed for updated beliefs to turn into behavioral changes. Our results suggest that future developments of Bayesian learning models should take this behavioral aspect into account. We have proposed a fairly simple, sequential model in which beliefs are first updated and any behavioral reactions afterwards depend on agents' motivation. But this may be more complicated in the sense that the motivational aspect may influence whether agents use new information at all, and if so, motivation may influence the way information is interpreted as suggested by the large literature on motivated reasoning (e.g., Christensen et al. 2018).

Furthermore, the motivation to react to performance information may also depend on whether information is provided at the level at which agents are able to act on it. Frontline workers may be more motivated to react on information about the specific clients they serve, rather than the organizations' average performance, whereas managers may be more motivated by information at the organizational level. More generally, accountability systems are designed to ensure that agents are motivated to react to the information. Very high-stakes accountability systems may create incentives for gaming the numbers and thereby create unintended consequences. Muller (2018) has recently argued that it is not the individual components per se, but the combination of measuring performance, publishing the results, and rewarding high-performance scores that creates negative effects.

Implications for Policy and Further Research

The most important implication of our study for policy is that policymakers should make sure students are tested regularly, such that systematic performance information is provided to coproducers such as principals, teachers, and parents. This seems to be particularly beneficial for low-SES students. In addition, our results indicate that student testing is valuable even in a context without strong economic incentives or threats of firing teachers and closing schools because of low performance.

One important caveat to our approach on studying components of performance management in isolation is, however, the concern that the effect of each component may depend on their interaction with each other. Whether perverse incentives in high-stakes accountability systems are stronger and crowd out the positive effects of learning from performance information is highly uncertain based on current evidence, though. Studies in higher-powered accountability systems in the United States and United Kingdom find effect sizes similar to those we find (Burgess et al. 2013; Figlio and Ladd 2015). Most existing studies examine the combined effects of performance management systems that are very complex in the sense that they combine many elements at the same time. If those studies produced consistently positive (or negative) effects, there would be less reason to examine key components separately. But because existing results are very mixed, we believe that the study of single components is promising for future research.

Another caveat is whether our results generalize to other policy areas and other countries. From a theoretical standpoint there is no reason to believe that (Bayesian) learning from performance information would be fundamentally different in other systems, but, as mentioned, the effect of generating performance information may interact with other components of performance management systems, and results may for that reason not directly generalize to different systems. Most studies of performance management have been conducted within education (Gerrish 2016), and more research from other policy areas is definitely needed. Again, our proposed approach of studying components of management systems separately is a fruitful way forward. Rather than randomly generating studies across the full universe of possible performance management systems across countries and policy areas, systematically testing the components and their interactions may be a more direct path to understanding the generalizability of the results and their scope conditions.

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

Supplementary material

Supplementary data are available at the *Journal of Public Administration Research and Theory* online.

References

Andersen, Simon Calmar. 2005. How to improve the outcome of state welfare services. Governance in a systems-theoretical perspective. *Public Administration* **83**:891–7.

Andersen, Simon Calmar. 2008. The impact of public management reforms on student performance. *Public Administration* **86**:541–58.

Andersen, Simon Calmar, Louise Voldby Beuchert, Helena Skyt Nielsen, and Mette Kjærgaard Thomsen. 2018a. The effect of teacher's aides in the classroom: Evidence from a randomized trial. *The Journal of the European Economic Association* forthcoming. doi:10.1093/jeea/jvy048

Andersen, Simon Calmar, Thorbjørn Sejr Guul, and Maria Humlum. 2018b. *Reducing the achievement gap between students of high and low socioeconomic status. Evidence from a field experiment*. Working Paper. Aarhus, Denmark: Aarhus Univ.

Andersen, Lotte Bøgh, Eskil Heinesen, and Lene Holm Pedersen. 2014. How does public service motivation among teachers affect student performance in schools? *Journal of Public Administration Research and Theory* **24**(3):651–71.

Andersen, Simon Calmar, and Helena Skyt Nielsen. 2016. Reading intervention with a growth mindset approach improves children's skills. *Proceedings of the National Academy of Sciences of the United States of America* **113**:12111–13.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden. 2007. Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics* **122**:1235–64.

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2018. *Testing*. NBER Working Paper No. 24836, Cambridge, MA: National Bureau of Economic Research.

Beuchert, Louise Voldby, Maria Knoth Humlum, Helena Skyt Nielsen, and Nina Smith. 2018. The short-term effects of school consolidation on student achievement: Evidence of disruption? *Economics of Education Review* **65**:31–47.

Beuchert, Louise Voldby, and Anne Brink Nandrup. 2018. The Danish national tests at a glance. *Danish Journal of Economics* **1**:1–37.

Black, Paul, and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* **5**:7–74.

Bond, Trevor G., and Christine M. Fox. 2007. *Applying the Rasch model. Fundamental measurement in the human sciences*, 2nd ed. New York, NY: Routledge.

Bourdieu, Pierre. 1984. *Distinction: A social critique of the judgement of taste*. London, UK: Routledge.

Boyne, George. 2010. Performance management: Does it work? In *Public management and performance: Research directions*, eds. Richard Walker, George Boyne, and Gene Brewer, 207–26. Cambridge, UK: Cambridge Univ. Press.

Burgess, Simon, Deborah Wilson, and Jack Worth. 2013. A natural experiment in school accountability: The impact of school performance information on student progress. *Journal of Public Economics* **106**:57–67.

Carnoy, Martin, and Susanna Loeb. 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis* **24**(4):305–31.

Chakrabarti, Rajashri. 2014. Incentives and responses under No Child Left Behind: Credible threats and the role of competition. *Journal of Public Economics* **110**:124–46.

Christensen, Julian, Casper Mondrup Dahlmann, Asbjørn Hovgaard Mathiasen, Donald P. Moynihan, and Niels Bjørn Grund Petersen. 2018. How do elected officials evaluate performance? Goal preferences, governance preferences, and the process of goal reprioritization. *Journal of Public Administration Research and Theory* **28**(2):197–211.

Danish Ministry for Children, Education and Gender Equality. 2011. *De nationale test og kommunen – Brug af testresultater i kommunens kvalitetsarbejde* (The municipality and the nationwide tests – Applying test scores in the municipal quality assessment). Copenhagen, Denmark: Ministry for Children, Education, and Gender Equality.

Dee, Thomas D., and Brian Jacob. 2011. The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management* **30**:418–46.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. School accountability, postsecondary attainment and earnings. *Review of Economics and Statistics* **98**:848–62.

Dizon-Ross, Rebecca. 2019. Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review* **109**(8):2728–65.

Figlio, David, and Helen Ladd. 2015. School accountability and student achievement. Ch. 12. In *Handbook of research in education finance and policy*, eds. Helen F. Ladd and Margaret E. Goertz, New York, NY: Routledge, 194–210.

Figlio, David, and Susanna Loeb. 2011. School accountability. Ch. 8. In *Handbooks of the economics of education*, eds. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, vol. **3**, Amsterdam, The Netherlands: Elsevier, 383–421.

Figlio, David, and Cecilia E. Rouse. 2006. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics* **90**:239–55.

Fitzpatrick, Maria D., David Grissmer, and Sarah 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**:269–79.

Gerrish, Ed. 2016. The impact of performance management on performance in public organizations: A meta-analysis. *Public Administration Review* **76**:48–66.

Hanushek, Eric A., and Margaret E. Raymond. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management* **24**:297–327.

Harrits, Gitte Sommer. 2019. Stereotypes in context: How and when do street-level bureaucrats use class stereotypes? *Public Administration Review* **79**:93–103.

Harrits, Gitte Sommer, and Marie Østergaard Møller. 2014. Prevention at the front line: How home nurses, pedagogues, and teachers transform public worry into decisions on special efforts. *Public Management Review* **16**(4):447–80.

Hattie, John. 2009. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Oxon, UK: Routledge.

Hattie, John, and Helen Timperley. 2007. The power of feedback. *Personality and Social Psychology Bulletin* **77**:1664–80.

Heinesen, Eskil. 2010. Estimating class-size effects using within-school variation in subject-specific classes. *Economic Journal* **120**:737–60.

Heinrich, Carolyn J., and Gerald Marschke. 2010. Incentives and their dynamics in public sector performance management systems. *Journal of Policy Analysis and Management* **29**:183–208.

Hess, Jacob, Henriette Holmsgaard, and Louise Weinrich Jaokobsen. 2009. Kvalitetsrapporten som kommunalt styringsredskab (In English: The quality assessment as an accountability measure). In *Kvalitetsrapporten: Evaluering og Udvikling*, ed. Tanja Miller. København, Denmark: Dafolo.

Holm, Jakob M. 2017. Double standards? How historical and political aspiration levels guide managerial performance information use. *Public Administration* **95**:1026–40.

Holm, Jakob M. 2018. Successful problem solvers? Managerial performance information use to improve low organizational performance. *Journal of Public Administration Research and Theory* **28**:303–20.

Hood, Christopher. 2011. *The Blame Game. Spin, Bureaucracy, and Self-Preservation in Government*. Princeton, NJ: Princeton University Press.

Jacob, Brian A. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics* **89**:761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* **118**(3):843–78.

Jakobsen, Morten L., Bækgaard, Martin, Moynihan, Donald P., and Loon, Nina van. 2018. Making sense of performance regimes: Rebalancing external accountability and internal learning. *Perspectives on Public Management and Governance* **1**:127–41.

Jakobsen, Morten, and Andersen, Simon Calmar. 2013. Coproduction and equity in public service delivery. *Public Administration Review* **7**:704–13.

Jussim, Lee, and Kent D. Harber. 2005. Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies. *Personality and Social Psychology Review* **9**:131–55.

Krieg, John M. 2011. Which students are left behind? The racial impacts of the No Child Left Behind Act. *Economics of Education Review* **30**:654–64.

Kroll, Alexander. 2013. The other type of performance information: Nonroutine feedback, its relevance and use. *Public Administration Review* **73**:265–76.

Lipsey, Mark, Kelly Puzio, Cathy Yun, Michael A. Hebert, Kasia Steinka-Fry, Mikel W. Cole, Megan Roberts, Karen S. Anthony, and Matthew D. Busick. 2012. *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: NCSER, US Department of Education.

Meier, Kenneth J., Simon Calmar Andersen, Laurence J. O'Toole, Natah Favero, and Søren C. Winter. 2015a. Taking managerial context seriously: Public management and performance in U.S. and Denmark schools. *International Public Management Journal* **18**(1):130–50.

Meier, Kenneth J., Nathan Favero, and Ling Zhu. 2015b. Performance gaps and managerial decisions: A Bayesian decision theory of managerial action. *Journal of Public Administration Research and Theory* **25**(4):1221–46.

Moynihan, Donald. 2008. *The dynamics of performance management: Constructing information and reform*. Washington, DC: Georgetown Univ. Press.

Muller, Jerry Z. 2018. *The tyranny of metrics*. Princeton, NJ: Princeton Univ. Press.

Neal, Derek, and Diane W. Schanzenbach. 2010. Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics* **92**:263–83.

O'Toole, Laurence J., and Kenneth J. Meier. 1999. Modeling the impact of public management: Implications of structural context. *Journal of Public Administration Research and Theory* **9**(4):505–26.

O'Toole, Laurence J., Jr., and Kenneth J. Meier. 2011. *Public management: Organizations, governance, and performance*. London, UK: Cambridge Univ. Press.

OECD. 2004. *OECD-rapport om grundskolen i Danmark*. Paris, France: OECD.

Petersen, Niels Bjørn G., Trine V. Laumann, and Morten Jakobsen. 2019. Acceptance or disapproval: Performance information in the eyes of public frontline employees. *Journal of Public Administration Research and Theory* **29**:110–17.

Rambøll. 2013. *Evaluering af de nationale test. (Evaluation of the national test)*. Aarhus, Denmark: Rambøll.

Rambøll. 2014. *Supplement til evaluering af de nationale test. (Supplement to evaluation of the national test)*. Aarhus, Denmark: Rambøll.

Reback, Randall. 2008. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics* **92**:1394–415.

Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review* **102**(7):3184–213.

Roediger, Henry L., III, Adam L. Putnam, and Megan A. Smith. 2011. Ten benefits of testing and their applications to educational practice. Ch. 1. In *Psychology of learning and motivation: Cognition in education*, eds. Jose P. Mestre and Brian H. Ross. Oxford, UK: Elsevier.

Rouse, Cecilia E., Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy* **5**:251–81.

Sherman, David K., and Geoffrey L. Cohen. 2006. The psychology of self-defense: Self-affirmation theory. *Advances in Experimental Social Psychology* **38**:183–242.

Snyder, Jeffrey W., Andrew Saultz, and Rebecca Jacobsen. 2017. Antipolitics and the hindrance of performance management in education. *Journal of Public Administration Research and Theory* **29**(4):e1–e7.

Tenenbaum, Harriet R., and Martin D. Ruck. 2007. Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology* **99**:253–73.