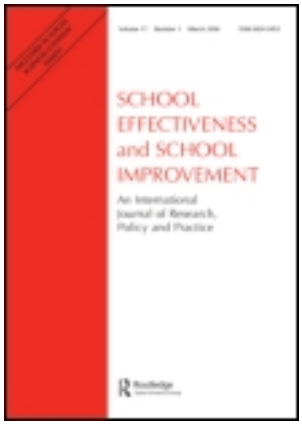
On: 11 April 2012, At: 01:55
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/nses20

Strategic data use of schools in accountability systems

Melanie C.M. Ehren [a] & Machteld S.L. Swanborn [b]

[a] Faculty of Behavioural Sciences, University of Twente, Enschede, The Netherlands

[b] Dutch Inspectorate of Education, Utrecht, The Netherlands

Available online: 04 Apr 2012

To cite this article: Melanie C.M. Ehren & Machteld S.L. Swanborn (2012): Strategic data use of schools in accountability systems, School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice, 23:2, 257-280

To link to this article: http://dx.doi.org/10.1080/09243453.2011.652127

School Effectiveness and School Improvement
Vol. 23, No. 2, June 2012, 257–280

Strategic data use of schools in accountability systems

Melanie C.M. Ehren[a]* and Machteld S.L. Swanborn[b]

[a]*Faculty of Behavioural Sciences, University of Twente, Enschede, The Netherlands;* [b]*Dutch Inspectorate of Education, Utrecht, The Netherlands*

School inspections are expected to have an impact on data use and improvement of schools. Schools are expected to generate data (e.g., self-evaluation reports and student achievement results) as part of the inspection process. This process, in turn, also generates data (e.g., inspection reports) for school improvement. The high-stakes context in which both types of data are generated however has been known to lead to strategic responses of schools. In this study, we analyzed if schools cheat on tests and reshape their test pool in responses to the Dutch (risk-based) school inspections. We found that 5.5% of the schools do not to comply with the guidelines for administering the test; one third of the schools exclude one or more students from the test. These responses, however, do not appear to be related to specific measures in the Dutch school inspections or prior performance of schools on these measures.

Keywords: data-driven decision making; school inspections; accountability; strategic behaviour

Introduction

Across the world, accountability systems have been introduced in education to monitor the performance of teachers and schools and to provide them with data to improve. Data are central to accountability systems. Schools generate data (such as student achievement data, data from self-evaluations, or parent and teacher surveys) to give an account on how they are doing; these data are used as part of the accountability system to assess performance of teachers and schools with respect to some externally defined target. In turn, this process generates data, such as school inspection reports or test scores of students compared to national targets, for schools to use in school improvement.

Schools are likely to feel substantial pressure to act in response to both types of accountability-related data because of the high-stakes context in which most data are generated. This context has been known to lead to strategic responses in schools that affect the types of data and the quality of the data teachers and schools use to inform teaching and improvement of the school.

Strategic responses, for example, include schools fixating on a small range of accountability measures and data used to evaluate their performance. Schools

*Corresponding author. Email: m.c.m.ehren@utwente.nl

ISSN 0924-3453 print/ISSN 1744-5124 online

http://dx.doi.org/10.1080/09243453.2011.652127
http://www.tandfonline.com

sometimes also distort data to improve their status on these accountability measures without creating a commensurate improvement in the educational processes or output the data are intended to measure. Teachers may, for example, correct answers of students on the test that is used to measure student achievement of the school. As a result, the data are inflated and lose their value to inform decision making of teachers and schools.

Strategic responses of teachers and schools are particularly likely when accountability systems base important decisions on a single measure of a limited number of aspects of teaching and learning and when high stakes are introduced for schools to perform well on these single measures. In the USA, many examples exist of schools and teachers using data to teach to the test as a result of high-stakes test-based accountability (see Booher-Jennings, 2005; Koretz, 2003). In Europe, a number of authors have explained how school inspections have caused schools to collect data for the inspectorate through surveys to specific groups of (satisfied) parents and teachers.

Several scholars have therefore expressed the need for multiple measures in accountability systems, to hold schools to account for broader goals and to prevent fixation of schools on a small number of quantifiable indicators (Barber, 2004; Koretz, 2003; Ladd, 2007). Multiple measures include instruments and methods of data collection to evaluate *cognitive outcomes, non-cognitive outcomes* (e.g., attendance and dropout rates), and/or *educational practices* in schools (such as teaching strategies) (Koretz, 2003). The use of multiple measures is expected to stimulate schools to use data to improve on a broader set of goals and to lessen incentives to engage in strategic behaviours that can distort data. Adding school inspections to test-based accountability may, for example, mitigate the extensive focus of schools on test results and redirect schools' attention to using data to improve other indicators of teaching and learning. On the other hand, the use of test results to measure output of schools in addition to school inspections can redirect schools' focus on producing data to describe their teaching and learning processes towards a focus on improving student achievement. Furthermore, multiple measures and indicators are more difficult to manipulate than just one single measure and indicator.

Currently, however, no studies are available that provide insight into whether these accountability models, using multiple measures, actually lessen the problems of strategic responses and increase data use for genuine school improvement. This study aims to contribute to this knowledge base by studying if and how multiple measures in the Dutch accountability system prevent or cause strategic behaviour in Dutch primary schools (students aged 4–12 years). These schools have to account for both cognitive outcomes and the quality of their educational practices. A change in inspection methodology in 2007, however, shifted the focus from measuring these two types of performance to the measure of cognitive outcomes; school inspections of educational practices are now only scheduled in some (high-risk) schools. This shift is expected to lead to an increase in strategic behaviour related to the cognitive outcome measures in the Dutch accountability system. In this article, we will analyze two types of potential strategic responses to measures of cognitive outcomes: teachers cheating on the test and schools reshaping their test pool to improve the school's performance on the test that is used by the Inspectorate of Education to measure cognitive outcomes. We will also analyze if low-performing schools (who face higher stakes to do well on the accountability measures) reshape their test pool more than high-performing schools.

The following research questions will be answered:

(1) To what extent do we find teachers and schools cheating on the test and reshaping their test pool in response to the accountability measures of cognitive outcomes in Dutch primary schools?
(2) Are these behaviours related to prior performance of schools and to a change in the number of measures in the accountability system?

The following section first presents a description of accountability systems using multiple measures. Next, we present potential strategic behaviour on measures of cognitive outcomes in Dutch schools. In the results section, we provide evidence of two types of these behaviours (cheating and reshaping the test pool). Finally, we try to explain how these behaviours are related to the measures in the accountability system and how they affect the use of data in schools.

Multiple measures in accountability systems

The term "multiple measure"' has been used in many ways. Andrews Paulsen, Ferrara, Birns, and Joffre Leclerc (2002) define multiple measures as more than one type of assessment or other achievement indicator (e.g., exit exams, course grades, portfolios, performance-based assessments) to measure student achievement. According to the American Institutes for Research (see Andrews Paulsen et al., 2002), multiple measures include both large-scale tests and other types of assessments in addition to course grades to measure student achievement. These definitions show that the concept of multiple measures is generally used to improve the reliability of measures of student achievement and student ability and to decrease the amount of measurement error in making decisions about student status. Baker (2003) and Gong and Hill (2001), however, also emphasize the use of multiple measures in accountability systems to increase the validity and reliability of judgements of schools. Multiple measures in this respect include different types of assessments or indicators to measure the educational quality of schools. In this article, we focus on the concept of multiple measures as the array of measures in an accountability system to evaluate a school's performance rather than the number of measures any particular student is expected to take.

These measures of performance of schools can be described on a number of different dimensions. We distinguish between the type of performance measured, the type of instruments (measure and sources of data) used, and the weighing of measures (through the setting of thresholds) used to reach a conclusion on performance of schools.

The first dimension, type of performance of schools, is described in detail by Koretz (2003). He distinguishes between distant, intermediate, and proximate cognitive outcomes, non-cognitive outcomes, and educational practices. Cognitive outcomes include achievement of students in different subjects. Accountability systems often focus on achievement of students in math and language, but some countries have also started to include achievement of students in other subject areas such as science or social studies. Koretz distinguishes tests to measure distant, intermediate, and proximate cognitive outcomes of schools. He describes distant outcomes as "aspects of achievement that are far removed in time from many of the immediate concerns of most teachers" (2003, p. 3). Most accountability systems use

tests that are carried out only in a handful of grades. For teachers in other grades, these outcomes are quite remote and compete for attention with more immediate goals, according to Koretz. Tests measuring intermediate outcomes are less far removed in time from the immediate concerns of teachers but are often tested with external assessments (Koretz, 2003). These tests are intermediate measures for the teachers teaching the pertinent classes, although they are distant measures (or even simply irrelevant) for other teachers. Proximate outcomes are the short- and moderate-term products of schooling that occupy much of the daily attention of many teachers (Koretz, 2003). According to Koretz, these tests (such as classroom-based assessments) are diverse and often measure increases in knowledge and skills and changes in motivation, other attitudes, and behaviour.

According to Baker (2003), non-cognitive measures traditionally include variables or indicators that have been treated in the literature as affective and psychomotor performance. Non-cognitive outcomes are often included in accountability systems to measure other key outcomes of teaching and learning that are worthy of monitoring. The U.S. Department of Education, for example, includes attendance, suspension, graduation rates, and dropout statistics as measures of non-cognitive outcomes of schools. According to Koretz (2003), these measures are sometimes used as a means of controlling inappropriate responses of schools to test-based accountability in the USA. Including these measures of non-cognitive outcomes may prevent schools from, for example, attempting to boost test scores by retaining students in a grade or by allowing low-performing students to drop out.

Measures of educational practices generally include inspection visits or quality reviews in which inspectors visit schools to evaluate educational practices related to teaching, organization, and leadership. These practices may, for example, include instructional and assessment practices, school climate, and/or methods used to accelerate student learning.

In addition to describing the type of performance measured in the accountability system, Lee and Coladarci (2002) and Gong and Hill (2001) describe an assessment dimension. This dimension refers to the type of instruments used to measure performance. The types of measures of cognitive outcomes in accountability systems generally consist of standardized external tests of students' performance, including multiple-choice, short-answer, or extended-constructed-response items. These tests represent intermediate or proximate outcomes for the teachers in the tested grades and distant outcomes for teachers in untested grades. In measuring cognitive outcomes of schools, there has also been discussion about including different types of assessments such as teacher assessments, local tests, and national standardized tests to measure one indicator or domain (Baker, 2003).

Instruments to measure non-cognitive outcomes often include schools reporting data on these indicators using records from their own management information systems. Educational practices are generally measured by means of school inspections or quality reviews in which inspectors observe lessons, analyze documents (such as self-evaluations of schools), and have meetings with teachers, the principal, parents, and students to gather information about the school's functioning on certain quality areas and standards. Some accountability systems also include self-evaluations of schools, peer reviews, or surveys. The Dutch Inspectorate of Education, for example, uses self-evaluations of schools as a source in their inspection of schools, while the accountability system of New York City uses surveys

to parents, students, and teachers to measure educational practices of schools such as their academic environment.

Gong and Hill (2001) and Chester (2003) point to the weighing of measures to come to judgements on the performance of schools. Data should be combined into a score or decision in a way that is useful, efficient, and defensible. Chester (2005), for example, describes a conjunctive, a compensatory, a confirmatory, and a complementary approach to combining multiple measures in accountability systems. Accountability systems using a conjunctive combination of measures require schools to attain standards on multiple measures; schools have to perform well on standards related to both tests and school inspections to be evaluated positively. Schools may compensate stronger results on the one measure with weaker results on the other measure in case of a compensatory combination of measures. Confirmatory combinations of measures use the results on one measure to validate the results on another measure.

Accountability systems measuring cognitive outcomes often include one of two types of models to weigh students' test scores on the outcome measures against pre-set thresholds (Gong & Hill, 2001; Hamilton & Koretz, 2002). The first type of model sets targets on desired levels of change in school performance. Schools reach acceptable levels of performance, according to the pre-set accountability threshold, when they show improvement in test scores each year. In a second type of model, test scores are weighed against pre-set thresholds which only report performance of schools (instead of change). These performance-reporting models may use norms to evaluate a school's performance in terms of its position in a distribution of scores of other (similar) schools (such as national percentile ranks or normal curve equivalents); or they may use criteria to evaluate a school's performance relative to a fixed level of performance (e.g., pass rates or test-score levels representing mastery of a specific range of content). Some accountability systems also use combinations of these thresholds; for example, they set targets on both minimum performance levels and improvement of student achievement each year.

Thresholds on non-cognitive outcomes may include maximum levels of dropout on tests or minimum attendance rates of students. A threshold to evaluate educational practices generally includes a selection of standards related to the quality areas in the overall inspection framework; an overall score of the school's performance is often calculated using a numeric system in which a school earns points for each standard. The final score reflects whether a school is failing or proficient.

Table 1 summarizes the type of performance/quality area, measures, and thresholds that may be combined in accountability systems using multiple measures.

In the next section, we will describe the measures used in the Dutch accountability system to evaluate and monitor performance of primary schools.

Describing the Dutch accountability system

The Dutch Inspectorate of Education is the main actor in monitoring performance of schools and in holding schools to account in The Netherlands. The Dutch Inspectorate used two different inspection methods in the period of our study. The first, "old", inspection method, enacted until 2007, included 3,450 short annual visits to all schools and full inspection visits of all schools every 4 years (on average 1,750 full inspections per year). Failing schools were visited more frequently. The second,

Table 1. Multiple measures in accountability systems.

Performance area	Type of measures and sources of data	Threshold
Cognitive outcomes: - Distant - Intermediate - Proximate	Standardized external tests, including multiple-choice, short-answer, or extended-constructed response items	- Improvement models including targets on change in student achievement, measured by means of a cross-sectional, quasi-longitudinal, or longitudinal approach - Performance-reporting model including targets on the position of a school in a distribution of scores of other schools, or fixed levels of performance
Non-cognitive outcomes: quantifiable indicators of dropout, attendance, graduation, students repeating a grade	Records of schools Self-evaluations Peer reviews	Cut scores of minimum/maximum acceptable levels on the indicators
Educational practices (teaching and learning, school climate, etc.)	School inspections (including interviews, lesson observations, document analysis)	Minimum score on a number of standards in the inspection framework

''new'', risk-based inspection method, implemented in 2007, includes early warning analyses in which student achievement results and other data sources are used to schedule inspection visits in a limited number of potentially failing schools (on average 615 visits per year). In addition, the Inspectorate makes sure that every primary school is visited at least once every 4 years with light-touch inspections in which only a small number of indicators of educational practices in the school are evaluated (on average 930 visits per year). This arrangement is in place to prevent schools without risks from having no inspection visit for many years.

Inspection methodology until 2007

Until 2007, Dutch school inspectors visited schools half a day every year for a short yearly inspection and once every 4 years for an extended quality inspection of 1 to 2 days. This extended visit was also carried out if the yearly visit showed defects in school quality. Before this visit, schools were requested to send information to the Inspectorate on their school policies. They also had to fill in questionnaires about, for example, their pedagogical vision, their lesson tables, the didactics they used, their instructional methods, and the test results of their pupils. During the inspection visit, the school's performance on the following quality areas was assessed:

(1) Did the students' results reach a level that may be expected (taking the characteristics of the student population into account)?

(2) Did the school have a system for assuring the quality of its education?
(3) Did the subject matter offered to pupils prepare them for continuing education?
(4) Did the students get enough lesson time to learn the subject matter?
(5) Did the school systematically assess the progress of students?
(6) Was the school climate safe and stimulating?
(7) Did the pedagogical behaviour of teachers meet basic requirements?
(8) Did the didactical behaviour of teachers meet basic requirements?
(9) Did children with specific educational needs receive the care they needed?

The Inspectorate used the results of students on the Cito test in Grade 8 to evaluate the first quality area, student achievement.[1] This quality area reflected the cognitive outcomes of the school, whereas quality areas 2 to 9 were related to educational practices in the school.

The Cito test, used to evaluate the cognitive outcomes of schools, is a norm-referenced standardized external test (administered by teachers), using multiple-choice items. The Cito test is administered in approximately 85% of primary schools. Schools are presented with an overview of the average performance of their students in general and on each subject compared to the relative average of schools with a similar student population.

The Cito test contains a compulsory part of 200 multiple-choice items on Dutch language, arithmetic, and study skills and a voluntary part of 90 items on history, geography, and science. The Cito test is based on the learning objectives that are specified for primary education by the Department of Education.

Schools that administer the Cito test are advised to have all students participate in the test, except for migrant students that have been in The Netherlands for less than four years at the start of the final Grade 8 in which the test is administered; these students do not have enough language skills to be able to understand the test items. Students that are expected to go to (advanced) special education may also be exempted from the test. Until 2007, a third category of students was also allowed not to take the test, namely, students that are expected to go to the learning support trajectory in regular secondary education.

Information on quality areas 2 to 9 was gathered in meetings with principals and teachers and by observation of lessons and analyses of school documents according to standard guidelines, protocols, and assessment forms (Janssens, Vanotterdijk, & De Wolf, 2005). Inspectors used the results of self-evaluations completed by the schools if they were considered to be reliable and valid; the self-evaluation results of schools had to provide data about the indicators included in the inspection framework. If these requirements were met, the Inspectorate would conduct fewer and less intense inspection visits.

The Inspectorate used a performance-reporting model to evaluate the school's performance on the cognitive outcomes and their educational practices. A primary school was identified as ''weak'' when:

- the student achievement results in Grade 8 were half a standard deviation below the relative national average of schools with a similar student population for at least three years; or when
- the student achievement results in half of the Grades 1 to 7 were half a standard deviation below the relative national average of schools with a similar

student population, and several indicators of the teaching-learning process were underdeveloped.

A school was identified as "highly underdeveloped" when the student achievement results in Grade 8 were below the relative national average of similar schools for at least three years and two or more of the standards of teaching and learning and pupil care were underdeveloped.

Schools were classified into one of seven groups of schools to calculate their score in relation to the national average of schools with a similar student population. Each group of schools has a student population with a similar socioeconomic status. This socioeconomic status is a construct based on the highest educational level of the pupils' parents and (until 2009) on their ethnic background. The school records the socioeconomic status of each student for this purpose. The school is failing on the measure of cognitive outcomes when the student achievement results on the Cito test at the end of primary education have been (more than half a standard deviation) below the average of schools in their group of schools.

Where the Inspectorate could not evaluate achievement levels of the school (e.g., in very small schools or schools that had not administered standardized achievement tests), the assessment of the quality assurance (standards related to quality area 2) was to assess if schools are failing.

Assessments of each school were described in an inspection report that was published on the internet to inform parents and the larger community of the school about the quality of the school. At the time of this study, school inspectors in The Netherlands have limited means with which to sanction schools. Sanctions (e.g., fines, closing schools) may be enacted only after a series of other measures have been taken (e.g., monitoring the school, replacing the school board). In addition, school inspectors may only recommend punitive measures to the Minister of Education; they may not enact such sanctions themselves.

Inspection methodology after 2007

Starting in 2007, the Dutch Inspectorate of Education implemented risk-based school inspections of schools. Every year, the Inspectorate carries out early warning analyses of all schools. In these analyses, information is collected on possible risks of low educational quality in all schools, such as student achievement results on standardized tests, self-evaluation reports and financial reports of schools, complaints of parents, and news items in the media. Results of students in Grade 8 of primary education (corrected for the socioeconomic background of students) on the national standardized Cito test are the primary indicator in the early warning analysis. The Inspectorate considers the student results to be a good predictor of the educational quality of schools on the nine indicators. It should be noted that the early warning analysis (using test scores) is not used to evaluate outcomes of the school. The test scores are only used as a first indicator of potential risks of low educational quality on the nine quality areas in schools. Both outcomes and educational practices are evaluated during additional inspection visits where the early warning analysis shows potential risks. If the early warning analysis shows no risks, schools receive no inspection visit during that year. The aim of these early warning analyses is to increase the efficiency of the Inspectorate by scheduling visits

only in schools that need it the most. The standards and thresholds used to assess whether a school is failing or well developed are unchanged.

Strategic data use

The change in inspection methodology implies that the measure of cognitive outcomes has become dominant in measuring school performance. In order to predict how schools may respond to this change in measure, this next section will present a literature review of potential strategic behaviour in schools related to measures of cognitive outcomes and the conditions in schools that promote these types of behaviour. We will also discuss potential positive and negative consequences of these types of behaviour.

Strategic behaviour related to measures of cognitive outcomes

Evidence of strategic behaviour of schools related to measures of cognitive outcomes can be found most prominently in the USA, where test-based accountability is the dominant form of accountability in education. Several studies give examples of how teachers and schools try to use data to influence test scores that are used to measure their output. Koretz, McCaffrey, and Hamilton (2001) identify a number of categories of responses to high-stakes testing and their likely effects on test scores and student learning. These responses may be positive when teachers and schools use data to improve teaching and learning, leading to valid increases in test scores, such as providing more instructional time or covering more material. Negative responses include misuse or cheating of data that leads to harmful consequences for student learning or inflated test scores.

Koretz et al. (2001), Stecher (2002), and Booher-Jennings (2005) also identify responses of teachers and schools and types of data use whose impact is ambiguous; the impact depends on the specific circumstances.

Ambiguous responses of teachers include schools and teachers using data or responding to accountability-related data to:

- reallocate instructional resources (classroom time or students' study time) to emphasize topics covered by the test instead of content that receives little or no emphasis on the test;
- coach students to do better by focusing instruction on incidental aspects of the test;
- align instruction with standards to give material and curriculum content that is consistent with standards more emphasis;
- shift teachers among grades to put less able teachers in untested grades;
- admit only students with advantaged backgrounds to the school;
- target instructional resources to students close to a cut-point set in the accountability system to improve the school's overall score on the accountability measures (educational triage).

These responses may have positive consequences when school principles and teachers use data to focus on important aspects of the domain the test is designed to measure or specific skills that help students demonstrate their actual achievement (Stecher, 2002); they may also have positive consequences for students in the tested

grades, the students who are admitted to the school, and students who are considered ''suitable for treatment''. Students in untested grades, students who are not admitted to the school, or students who are considered ''lost cases'' will, however, experience negative consequences.

Negative responses occur, according to Stecher (2002), when teachers respond to high-stakes testing by cheating and distorting data used to measure the school's status on the accountability measures. Jacob and Levitt (2003) revealed that cheating occurred in 4–5% of the classrooms each year. Teachers may do so by prompting students with the right answer during a test, by providing the actual test items in advance, by providing hints during test administration, by making changes to answer sheets before scoring, or by leaving pertinent materials in view during the testing session.

Figlio and Getzler (2002) and Cullen and Reback (2006) also describe how schools at risk of failing improve their state-assigned grade or classification by taking their poorest performing students out of the testing pool. This type of strategic behaviour is usually referred to as ''reshaping the test pool''. Schools may do so by reclassifying (regular) students into the ''special education'' or ''limited English proficient'' categories that may be exempted from taking the test (Jacob, 2005). Other methods used are retaining low-scoring students in grades below those in which the test is administered, allowing an increase in absences on test days, granting requests for exemptions from testing by parents of low-achieving students, and increasing dropout rates of low-achieving students. Cheating and reshaping the test pool may lead to inflation of scores. As a result, teachers have no test data to inform their instruction of certain groups of students, and the test data in general lose their value to inform decision-making of teachers and schools.

Relevant conditions for strategic behaviour of schools

Strategic behaviour will probably only occur in some schools in some conditions. Schools performing well will not be inclined to use strategic behaviour to meet performance targets. Hanushek and Raymond (2002), for example, expect schools that have scores close to a performance target to alter their behaviour more than schools further away from the established performance target. The interrelationship between the choice of school score model, the choice of performance targets, and the location of a given school relative to those targets will influence strategic behaviour in schools. De Wolf and Janssens (2007) also argue that low-performing schools and schools with a large population of minority pupils will be more inclined to use strategic behaviour to meet accountability standards. Jacob and Levitt (2003) and Stecher (2002), for example, found that teachers in schools with lower achievement, higher poverty rates, and more Black students were more likely to cheat and engage in excessive test preparation more frequently.

Potential strategic data use in Dutch primary schools

In this section, we will use the examples of strategic data use in the previous section to predict potential strategic behaviour of Dutch primary schools on the inspection measures and thresholds on cognitive outcomes. In sum, after the introduction of risk-based school inspections in 2007, the Inspectorate uses the following measures to evaluate the cognitive outcomes of schools:

(1) Student achievement results at the end of schooling are measured through the Cito test.
(2) The Cito test measures proximate outcomes for the principal and for teachers in Grade 8; other teachers will perceive the measure as intermediate or distant.
(3) The Cito test measures Dutch language, arithmetic, and study skills and includes a voluntary part in history, geography, and science.
(4) Results of all students in Grade 8 are used to assess outcomes of schools; migrant students and students who are expected to go to special education and students that will receive extra learning support in secondary education may be exempted from the test.
(5) A school is failing when the student achievement results on the Cito test at the end of primary education have been (more than half a standard deviation) below the relative national average for at least three years, compared to schools with a similar student population.

We expect teachers in Grades 7 and 8 and administrators in schools that score below the relative national average of similar schools in the last one or two out of three years on the Cito test in Grade 8 to use strategic behaviour to boost their test scores and improve the school's status on the outcome measures. These schools are in danger of performing below the threshold; the Inspectorate will assess them to be failing if next year's students' results are below average. These teachers and administrators may try to improve their student outcomes and/or their status on the outcome measures by means of *providing more instructional time in language and arithmetic, working harder*, or *more effectively covering more material*. These schools may also choose to *reallocate* classroom instructional time to language and arithmetic and away from untested subjects such as history, geography, and science. Schools are also expected to reallocate instructional resources and align the curriculum and instructional materials to specific topics within Dutch language, arithmetic, and study skills such as writing skills and grammar, as these topics are always part of the test. They may ignore other topics (such as oral language skills) that are never part of the test. Schools may also choose not to participate in the voluntary part of the test, which includes items on history, geography, and science, to mask their lack of teaching in these subjects.

Teachers in Grade 7 and particularly Grade 8 are also expected to *coach* their students to do better on incidental aspects of the test. They may teach their students to select the correct answer in multiple-choice items; they may practice old tests as part of classroom instructions, or they may use other coaching materials to try and improve students' scores on the test.

Administrators in schools are expected to *shift teachers* with low language or arithmetic skills to untested grades (Grades 1 to 6). As Grades 7 and 8 are closest to the administration of the test, schools are expected to put their best teachers in these grades.

The guidelines on exclusion of students from the Cito test provide schools with an opportunity to *reshape the test pool*. Teachers in Grade 8 and administrators may try to boost the school's status on the outcome measure by classifying students as going to special education, excluding migrant students, and stimulating absence of disadvantaged students on the test day. As the Inspectorate of Education compares schools to the average student achievement results of other similar schools with

respect to the number of disadvantaged students, this type of behaviour will be beneficial for schools to improve their status on the outcome measure of the Inspectorate. Schools may also retain students in Grade 7 to try and improve their scores on the test. Schools will choose to retain these students in Grade 7 and not in a lower grade because the Inspectorate uses the Cito scores of the last three years in their outcome measure of schools; schools have to perform at or above average at least one out of three years. They will choose to have students repeat a grade when their average score in the last two years was below average. These schools are in danger of performing below the inspection threshold and will feel the need to improve next year's average score quickly by retaining some students in Grade 7 to provide them with an extra year of teaching. The last option schools may choose to reshape their test pool is by referring low-achieving students to special primary education or (advanced) special education during their primary school career and before Grade 8. These students will not be part of the test pool in Grade 8 and as a result cannot lower the school's status on the outcome measure.

As the Cito test is administered by teachers to their own students, teachers may also *cheat* by providing their students with correct answers, allowing students to use supporting materials when making the test (e.g., books, examples, calculators, etc.), or correcting students' answer sheets.

Methods

In this article, we focus on the occurrence of schools cheating in administering the Cito test and reshaping their test pool. Unfortunately, we do not have information on other types of strategic behaviour in primary schools, and we also only have data for one point in time on potential cheating of schools. The available data allow us to analyze if schools reshape their test pool and cheat in administering the test and whether the occurrence of reshaping of the test pool increased after the introduction of risk-based school inspections. In the following sections, we will describe the two data files that were used to measure potential cheating and reshaping of the test pool.

Data on cheating during administration of the Cito test

The Dutch Inspectorate measured the extent to which teachers cheat in administering the Cito test through observation of teachers during the administration of the Cito test in Grade 8 in primary education in 2006. School inspectors visited 257 primary schools (3.7% of all primary schools; 4.3% of the primary schools who administer the Cito test) during the day the Cito test was scheduled to be administrated to students in Grade 8; 183 of these school visits were not announced to schools. School inspectors observed whether the test was administered according to the guidelines of the testing company.

Schools were selected that were either scheduled for an inspection visit during the time frame of the test administration or that classified an above-average number of students as moving onto the learning support trajectory in secondary education. This second category of schools is suspected of acting strategically during test administration. The selection of schools may affect our results in pushing up the number of schools we find to cheat during test administration. The fact that school inspectors are present during test administration will, on the

other hand, limit the number of cheating schools and teachers. A comparison of our sample of schools to the entire population of primary schools also showed that our sample is representative of the wider population in terms of school size, socioeconomic background of students, regional dispersion, denomination, and educational quality. Table 2 presents an overview of the schools observed by school inspectors.

Table 2. Sample of schools to measure cheating during test administration.

	Sample of schools
School size	
average number of students	221.18
0–100 students	12.8% (36 schools)
101–200 students	34.0% (96 schools)
201–400 students	47.9% (135 schools)
More than 400 students	5.3% (15 schools)
Socioeconomic background of students	
>50% students in category 0 (no disadvantaged socioeconomic background)	75.5% (213 schools)
>50% students in categories 0.25 and 0.90 (minor to substantial disadvantaged socioeconomic and ethnic background)	7.8% (22 schools)
>50% students in category 0.25 (minor disadvantaged socioeconomic background)	1.4% (4 schools)
>50% students in category 0.90 (substantial disadvantaged socioeconomic and ethnic background)	15.2% (43 schools)
Regional dispersion	
Not located in a city	65.2% (184 schools)
Located in a small city (G32)	24.5% (69 schools)
Located in a large city (G4)	10.3% (29 schools)
Denomination	
Public	37.9% (107 schools)
Catholic	30.5% (86 schools)
Protestant	27% (76 schools)
Other	4.6% (16 schools)
Educational quality	
Weak (Inspectorate assigned increased monitoring trajectory to school)	0.7% (2 schools)
Below average (Inspectorate assigned accelerated follow-up visit)	9.2% (26 schools)
Average or good (Inspectorate assigned regular inspection visit)	51.4% (154 schools)
Unknown	38.7% (109 schools)

Data on schools reshaping their test pool

The extent to which schools reshape their test pool was measured in all primary schools in The Netherlands from 2004 to 2009. We analyzed which students were exempted from the Cito test in Grade 8 and whether schools retained pupils in Grade 7 (prior to test taking) or referred students to special primary education or to (advanced) special education before Grade 8.

A comparison of the Cito database, school records, and school responses on the (annual) inspection questionnaires was used to compare the number of students in the eighth grade to the sum of the number of students that took the test and the number of students that where registered as not taking the test. This comparison shows how many students were exempted from the test. In the annual inspection questionnaires, schools reported why students were exempted from the test. The questionnaires also record the number of students retained in Grade 7 and the number of students referred to special primary education and (advanced) special education in the grades before Grade 8.

In order to examine reshaping of the test pool in the period 2004–2006, all 3,822 schools were selected that received the annual questionnaire in March 2006. In that time period, about half of the schools received the questionnaire in March, while the remaining schools received the questionnaire in September. This split into two tranches was based on the schedule of the inspection visits. There was no specific selection criterion on placing schools in one of the two groups. Schools that did not administer the Cito test were filtered out, as were schools where responses to the questions on the number of participating and non-participating students in the Cito test did not add up to the total number of students in Grade 8; 2,929 schools remained in our sample. In the questionnaire, schools were asked to provide participation rates of (different groups of) students on the national test and potential reasons for students not taking the test (e.g., sick leave) for the three consecutive years 2004–2006.

In the period after 2007, there is no longer an annual questionnaire for all schools. However, schools that are identified as being at risk in the early warning analysis and schools that are part of a national improvement program in language and mathematics still receive questionnaires that include similar questions as the former annual questionnaire. The schools that are part of the national language and mathematics program, however, vary widely in school quality. We therefore drew a representative sample of schools ($n = 1{,}184$) from this group to investigate whether schools reshape their test pool. The sampling was based on school size, composition of school population, regional dispersion, denomination, and educational quality. In the questionnaire, schools answered questions about the test pool in February 2009, about retaining pupils in Grade 7, and about referral of students to types of special education in the school year 2008–2009.

Table 3 provides an overview of the schools selected to measure ''reshaping of the test pool''. The table shows that the groups of selected schools before and after 2007 are relatively similar in school size, socioeconomic background of students, and denomination. The fact that the questionnaire was administered to all schools before 2007 and only to a selection of schools receiving an inspection visit after 2007 results in differences in information that is available on the educational quality of the selected schools. The selection of schools before 2007 includes a relatively large group of schools with ''unknown educational quality'', while the selection of schools after 2007 include no schools with ''unknown educational quality'' and a larger

Table 3. Sample of schools to measure "reshaping of the test pool".

	Sample 2004–2006	Sample 2009	Population 2009
School size			
average number of students	215	217	227
0–100 students	20.2%	19.1%	17.1%
101–200 students	31.5%	34.1%	31.2%
201–400 students	39.1%	38.1%	41.2%
More than 400 students	9.2%	8.7%	10.6%
Socioeconomic background of students			
>50% students in category 0 (no disadvantaged socioeconomic background)	88.3%	84.8%	85.5%
>50% students in categories .25 and .90 (minor to substantial disadvantaged socioeconomic background)	4.4%	5.7%	5.3%
>50% students in category 0.25 (minor disadvantaged socioeconomic background)	0.6%	1.9%	1.3%
>50% students in category 0.90 (substantial disadvantaged socioeconomic background)	6.7%	7.5%	7.8%
Regional dispersion			
Not located in a city	75.7%	71.6%	72.2%
Located in a small city (G32)	16.3%	20.0%	18.8%
Located in a large city (G4)	7.5%	8.4%	9.0%
Denomination			
Public	36.3%	33.7%	33.8%
Catholic	26.4%	30.2%	29.3%
Protestant	30.7%	30.7%	29.7%
Other	6.6%	5.4%	7.2%
Educational quality			
Weak (Inspectorate assigned increased monitoring trajectory to school)	0.8%	1.2%	1.5%
Below average (Inspectorate assigned accelerated follow-up visit)	7.9%	6.4%	7.4%
Average or good (Inspectorate assigned regular inspection visit)	57.7%	92.4%	91.1%
Unknown	66.4%	–	–

*Note: socioeconomic background in 2009 is only considered for pupils in the age 11–14. The measure of socioeconomic background changed in a way that a comparison to the period 2004–2006 could only be made for this age group.

group of schools with average or good educational quality. We do not, however, expect these differences to affect our results as the group of schools with "unknown educational quality" will also include a large group of schools with average or good educational quality.

The occurrence of reshaping of the test pool under the old and new inspection regime (before and after 2007) will be compared (using a *t* test and an analysis of

variance) to analyze whether the shift from multiple measures of cognitive outcomes and educational practices in all schools (before 2007) to a focus on measures of cognitive outcomes (after 2007) has led to an increase in this type of behaviour in schools. We expect an increase in schools reshaping their test pool as schools may be motivated to (only) generate and use data related to the measures of cognitive outcomes.

A second comparison includes schools that differ in their educational quality. A *t* test and an analysis of variance were used to compare reshaping of the test pool in schools that were identified by the Inspectorate as weak or underperforming to schools that perform well. Underperforming schools are expected to reshape their test pool to increase their status on the outcome measures, while high-performing schools are expected to refrain from reshaping their test pool.

A third comparison involves schools at risk. Since 2007, schools that have insufficient student achievement results for the past two years are considered to be at risk. If next years' student achievement results are below average, they will perform below the inspection threshold and will face increased monitoring. These schools are also expected to reshape their test pool to increase their status on the outcome measures, while high-performing schools are expected to refrain from this type of strategic behaviour.

Results

In this section, we present the extent to which schools deviate from test administration guidelines and potentially cheat during test administration and reshape their test pool. When analyzing the numbers of schools that reshape their test pool, we compare low- and high-performing schools and the numbers of schools reshaping their test pool before and after the change in inspection methodology in 2007. The numbers of schools that cheat during test administration only apply to 2006 (before the introduction of the new risk-based inspection methodology).

Cheating during test administration

School inspectors observed teachers during the administration of the Cito test. They found that 5.5% of the schools (14 schools) did not comply with the guidelines on administration of the test. The teachers in these schools allowed students to use scrap paper while completing the test; they clarified test questions or prompted students with the correct answer. Some teachers also gave instruction to students just before administration of the test (when having read the test items) or put up explanations of spelling problems on the blackboard. One of the schools also administered the test one day before the official test administration day.

Reshaping the test pool: exempting students from the test

Schools may reshape their test pool by exempting students (who are expected to perform below average) from the test. In Table 4, the percentage of schools that do not include all the students in the Cito test is presented over a period of 3 years. The table also indicates why students did not participate in the test. These motives and numbers were taken from the annual inspection questionnaires schools had to complete.

Table 4. Percentage of schools excluding students from testing.

	2004 ($n = 2{,}873$)	2005 ($n = 2{,}902$)	2006 ($n = 2{,}929$)	2009 ($n = 869$)
Schools exempting students	25.6	29.0	32.6	30.7
Motives to exempt students from testing:				
- Students absent due to illness	3.3	5.3	8.6	–
- Students referred to special education	6.0	7.4	8.3	–
- Students will receive extra learning support in secondary education	14.8	18.1	22.9	–

The results in Table 4 indicate that in 2006 almost one out of every three schools had at least one student who did not participate in the test. Over the years 2004 until 2006, the percentage of schools with students who did not take the Cito test grew steadily. The motives schools reported for absence of these students during the test were illness of students, referral to special education, or referral to the learning support trajectory in secondary education. The percentage of schools that reported student absences due to illness tripled in 2 years. Schools mostly excluded students who will be referred to the learning support trajectory in secondary education from the test; these percentages also increased over the years.

A paired sample t test was used to analyze if there were significant differences between the years in the number of students not participating in the test. These analyses show significant differences between the number of students exempted from the test from 2004 to 2005 and from 2005 tot 2006 ($p < .00$). An independent t test analysis between 2006 and 2009 showed no difference between the percentage of schools excluding students ($t = -1.048$, $p = .295$).

The stable numbers of schools that exempted students from the test in 2006 and 2009 is remarkable considering the fact that regulations on exclusion of students from the Cito test changed in 2007. Starting in 2007, schools were required to include all students who will receive extra learning support in regular secondary education in test administration. This requirement should have led to a significant decrease in the percentage of exempted students after 2007.

Next, we analyzed potential differences in the extent to which low- and high-performing schools exempted students from the test. The results in Table 5 provide an overview of the average percentages of students per school who did not participate in the Cito test from 2004 to 2009. A distinction is made between underperforming and well-performing schools. The results indicate that in each school on average 3 to 5% of the students did not participate in the Cito test.

An analysis of variance was performed to learn if underperforming schools excluded students from the test more frequently than well-performing schools. The results of this analysis show differences between underperforming and well-performing groups in 2005 ($F = 5.530$, $p = 0.019$). We found no significant differences between the two groups of schools in the years 2004, 2006, and 2009.

We also performed a paired-sample t test to examine whether the percentage of students not participating in the test rose over the years in all schools, and

Table 5. Percentage of students per school exempted from test-taking in underperforming and well-performing schools.

	2004 (N schools = 2,873) (N students = 71,396)	2005 (N schools = 2,902) (N students = 71,020)	2006 (N schools = 2,929) (N students = 71,466)	2009 (N schools = 869) (N students = 25,986)
Underperforming schools	3.8	5.0	4.4	4.7
Well-performing schools	3.1	3.4	3.8	3.2
Total	3.1	3.6	3.8	3.3

specifically in underperforming and well-performing schools. The results point to a significant rise in non-participation of students in all schools from 2004 to 2005 ($t = -3.721$, $p = 0.00$), but not from 2005 to 2006. If the group of underperforming schools is taken separately, there is a significant rise in exclusion from 2004 to 2005 only at the $p < 0.1$ level. The well-performing schools show significant changes in participation of students in the test from 2004 to 2005 and from 2005 to 2006, only at the $p < 0.1$ level. We found no significant differences in the number of non-participating students in well-performing schools between 2006 and 2009.

The last method was a comparison of the number of schools at risk to other schools on the extent to which they exempted students from test taking. Schools at risk have had below-average student achievement results for the past 2 years. As they are at risk of performing below the threshold the next year, they are expected to reshape their test pool. The results, however, show no significant differences between schools at risk and other schools. About 35% of the schools at risk were found to have students not taking the test, while about 30% of the schools that are not at risk have students not taking the test.

Reshaping the test pool: retaining students in Grade 7

We also examined the extent to which schools reshaped their test pool by retaining low-achieving students in Grade 7 prior to the year of test taking. Data were only available for the school years 2004–2005 and 2007–2008. The results in Table 6 indicate only small percentages (ranging from 0.5 to 1%) of students per school who were retained in Grade 7. These percentages were relatively stable over the years and did not increase after the introduction of risk-based school inspections in 2007.

We also used a means analysis to compare the percentage of students retained in Grade 7 in low-performing schools and high-performing schools. In 2005, there was no difference between these groups; in 2008, a difference existed at the $p < 0.1$ level between well-performing and underperforming schools; well-performing schools retained more students in Grade 7.

These differences, however, do not hold when we perform this analysis on schools at risk; we found no significant difference between groups of schools that are at risk and that are not at risk in the extent to which they retain students in Grade 7 ($F = 0.110$; $p = 0.740$).

Reshaping the test pool: referring students to special primary education or to (advanced) special education

Schools may also reshape their test pool by referring students to types of special education prior to Grade 8. These students are not part of the student population in Grade 8 anymore and as a result do not have to participate in the Cito test. The results in Table 7 provide an overview of the number and percentages of students referred to special education in all schools and in underperforming and well-performing schools. The average percentage of students per school that is referred to special primary education before Grade 8 is 0.2 to 0.3%.

An analysis of variance shows that underperforming schools referred significantly more students to special primary education than well-performing schools in 2007–2008. However, these differences do not hold when we perform this analysis on schools at risk; no significant difference was found between groups of schools that are at risk and that are not at risk in the extent to which they refer students to special primary education or special education ($F = .960$; $p = 327$, $F = .254$, $p = .614$).

Underperforming schools were compared to well-performing schools in 2004–2005. In that period, fewer students were referred to special (advanced) education. Analysis of variance showed no difference between underperforming and

Table 6. Percentage of students per school retained in grade 7.

		Students retained in Grade 7	
		2004–2005	2007–2008
Underperforming schools	Average %	0.8	1.0
	SD	2.4	3.0
	N	175	81
Well-performing schools	Average %	0.7	0.5
	SD	3.1	2.0
	N	1,645	1,038
Total	Average %	0.7	0.6
	SD	3.0	2.1
	N	1,820	1,119

Table 7. Percentage of students per school referred to special primary education or to special (advanced) education before Grade 8.

		Students referred to special primary education*	Primary school students referred to special (advanced) education
Underperforming schools	Average %	0.6	0.2
	SD	2.3	1.1
	N	80	80
Well-performing schools	Average %	0.1	0.3
	SD	1.0	1.7
	N	1,026	1,027
Total	Average %	0.2	0.3
	SD	1.2	1.7
	N	1,106	1,107

Note: *significant at $p < 0.05$.

well-performing schools for both types of referral ($F = 1.783$, $p = .182$; $F = 0.490$, $p = .484$).

Discussion

The research presented here has a number of limitations that need to be taken into account before presenting our conclusions. First of all, we were only able to include two out of the range of potential responses of teachers and schools to the use of data in (high-stakes) accountability systems. These two responses, cheating and reshaping of the test pool, are, however, very relevant to data use in schools. Both types of responses may inflate test scores and may cause teachers and schools to use biased data to inform their decisions on teaching and improvement of the school. We therefore consider these two responses as very relevant to studying how accountability measures affect data use in schools.

There are, however, some drawbacks in our collection of data on both types of responses. Potential cheating of teachers in the administration of the national standardized test was measured through school inspectors observing teachers while administering the test. The fact that school inspectors were in the classroom during test administration could have affected the number of incidences of cheating. We expect the number of incidences we found in our study to be lower than what the number would have been if there was no school inspector present.

Another drawback in our data collection is the fact that we only measured potential incidences of schools reshaping their test pool by means of comparing reported numbers of tested students in different categories. During our study, the policy of the testing company on inclusion of students in the test changed; schools were required to include an additional category of students in the test (students that will be referred to the learning trajectory in secondary education). This change in testing policy was introduced at the same time as the Inspectorate introduced their new inspection regime, which was expected to lead to an increase in exclusion of students from the test in our study. The net effect of both occurrences may have resulted in finding no change in the extent to which schools reshape their test pool through exempting students from the test. We, however, also measured two other examples of reshaping of the test pool (through retaining students in the grade prior to the testing grade and through referring students to special education) that are not affected by this change in guidance. A comparison of low- and high-performing schools provided insight into how the accountability measures affect schools in reshaping their test pool. The lack of a control group, however, limits us in making strong causal inferences on potential strategic data use resulting from (high-stakes) accountability systems.

The results of our study therefore primarily serve as a first indication of potential strategic responses to school inspection systems using measures of cognitive outcomes of schools. As there was no research available in this area until now, we consider our analysis of both types of responses to be a valuable contribution to the current research on data use and accountability, particularly given the fact that these responses have only been studied in test-based accountability systems.

Conclusion

Many accountability systems aim to stimulate and support schools in generating data for both accountability and improvement purposes. The high-stakes context in

which data are generated has, however, been known to lead to strategic responses in schools, which limits the combination of both goals as part of one accountability system. This seems to be particularly the case when schools are held accountable for their performance on a small number of quantifiable indicators. As a result, several scholars have expressed the need for more balanced accountability systems in which schools have to generate data related to multiple measurements and indicators of their performance. The use of multiple measures (such as measures of cognitive outcomes, non-cognitive outcomes, and educational practices) is expected to stimulate schools to use data to improve on a broader set of goals and to lessen incentives to engage in strategic behaviours that can distort data.

In this study, we analyzed two types of potential strategic responses (cheating and reshaping of the test pool) in Dutch primary schools related to the Dutch accountability system. The Dutch accountability system includes both measures of cognitive outcomes (primarily student achievement results on the so-called Cito test in Grade 8) and school inspections of educational practices to evaluate schools. A change in inspection methodology in 2007, however, shifted the focus to cognitive outcomes as school inspections are only scheduled in some (high-risk) schools that have relatively low cognitive student outcomes. As a result, we expected an increase in cheating and schools reshaping their test pool after 2007. We also expected low-performing schools to act more strategically compared to well-performing schools as they will face a higher stake to use strategic behaviour to meet the accountability standards.

We used two data files containing information on the extent to which teachers adhered to testing guidelines (and potentially cheated in administering the Cito test) in 2006 and containing information on the students that did not take the Cito test, that were retained in Grade 7, or were referred to special education before test taking between 2004 and 2009 (pointing to potential reshaping of the test pool). The Dutch Inspectorate measured the extent to which teachers ignored testing guidelines in administering the Cito test through observation of teachers during the administration of the Cito test in Grade 8 in primary education in 2006. The extent to which schools potentially reshape their test pool was measured in all primary schools in The Netherlands from 2004 to 2009. A comparison of the Cito database, school records, and school responses on the (annual) inspection questionnaires was used to analyze which students were exempted from the Cito test in Grade 8 and how many students were retained in Grade 7 (prior to test taking) or referred to special (primary) education before Grade 8.

The results of our study indicate that 5.5% of the schools do not comply with the guidelines on administration of the test. The teachers in these schools allowed students to use scrap paper while making the test; they clarified test questions or prompted students with the correct answer. Some teachers also gave instruction to students just before administration of the test (when having read the test items) or put up explanations of spelling problems on the black board. One of the schools administered the test one day before the official test administration day. These results date back to 2006, when schools still had to account both for their performance on cognitive outcomes and for their educational practices. We can make no comparison to the new inspection methodology, which primarily focuses on measures of cognitive outcomes. These results are therefore difficult to attribute to the measures in the Dutch accountability system. However, results of other studies (see Jacob & Levitt, 2003) also indicate 4–5% of teachers cheating or incorrectly administering

student achievement tests in single-measure (test-based) accountability systems. We therefore expect this percentage to represent a relatively fixed number of schools using such behaviour in any type of accountability system using measures of cognitive outcomes, unrelated to the specific weighing or consequences related to this measure.

One major drawback to this conclusion and to our results, however, is the fact that we did not ask schools for their motives and arguments to deviate from the Cito guidelines when administering the test or when taking students out of the test pool. As a result, we do not know if schools took these students out of the test pool to improve their status on the outcome measures or whether they had other motives. The fact that inspectors were present while teachers allowed students to use scrap paper during test taking or prompted students with the correct answer to test questions points to other motives for deviating from the test guidelines. Potential explanations may be that teachers are not aware of these guidelines or feel they are acting in the best interest of their students.

Our results also indicate that schools reshape their test pool to some extent, particularly by taking students out of the test pool that will be referred to regular secondary education with special learning support. Over the years of our study, at least one student did not participate in the Cito test in one out of every three to four schools. On average, in each primary school 3 to 5% of the students did not participate in the Cito test in Grade 8. The percentage of schools with students that did not take the Cito test also grew steadily over the years but seemed to stabilize after 2006. This may be related to a change in the testing guidelines issued by the testing company (Cito). From 2007 onwards, schools were required to include all students who will be referred to the learning support trajectory in secondary education in the test. This change in testing guidelines may have counterbalanced a potential increase in schools reshaping their test pool resulting from the introduction of risk-based school inspections in 2007. The lack of convincing differences in the number of excluded students in high- and low-performing schools, however, seems to point to other motives for these responses than to improve the inspection evaluation. Low-performing schools would have shown larger numbers of students excluded from the test if this response was motivated by the inspection measure of cognitive outcomes.

There are also few indications of schools using other methods to reshape their test pool; the percentages of students who were retained in Grade 7 or who were referred to special primary education before Grade 8 were very small. Schools most likely retained their students in Grade 7 or referred them to special primary education before Grade 8 because of actual low achievement of these students and the need to provide these students with extra or adjusted teaching and education (instead of doing so to increase the school's score on the outcome measure of the Inspectorate).

The results of our study nevertheless point to examples of schools who adjust their data, particularly in not adhering to testing guidelines and excluding one specific group of students (those that will be referred to the learning trajectory in secondary education). These responses may bias data and may affect the decisions that are based on these data.

Not adhering to the testing guidelines may, for example, bias the inspection measure of the school's performance to some extent. The Inspectorate of Education may assess the school as performing adequately, while performance is actually failing

or at risk. Also, principals may consider their school to be performing well and refrain from making changes in, for example, curriculum or teaching. However, we only found a small amount of schools not adhering to the testing guidelines, and the examples of ignoring the testing guidelines are not expected to cause serious bias to the test scores.

Reshaping of the test pool will, however, affect the information teachers have on the performance of specific groups of students and the extent to which the test scores represent performance of these students in the school. Teachers, for example, may have no information on students that will be referred to the learning trajectory in secondary education and cannot use their test scores in giving them advice on placement in specific levels of secondary education. If schools reshape their test pool for a longer period of time, they may also fail to have information on how these students are actually performing in the final grade of elementary education. As a result, teachers will not have information to potentially adapt their instruction to these students, and schools will miss an opportunity to evaluate whether their curriculum and teaching sufficiently facilitates learning of these students.

So far, the results of our study point to limited cheating and reshaping of the test pool, both before and after the introduction of risk-based school inspections, and in low- and high-performing schools. This seems to indicate that a change in the number of accountability measures does not affect strategic data use of schools. Perhaps schools will turn to such behaviour after a longer period of time (when they have become more aware of the increased focus on test scores in the inspection methodology) or when the stakes to meet certain scores are higher. Additional research may shed more light on this issue.

Note

1. Schools are allowed to choose a test to evaluate student achievement. The Inspectorate has developed guidelines on how to use results of each of the available tests to evaluate cognitive outcomes of schools. Most schools (85%), however, use the Cito test, administered by the testing company Cito. In this section, we will therefore only report on potential strategic behaviour of schools using the Cito test. The Inspectorate also uses Dutch language and arithmetic tests in Grades 3, 4, and 6 that are part of the Cito pupil monitoring system to evaluate progress of students during their school career (as a subindicator of the first quality area on cognitive outcomes). As the Cito test in Grade 8 is the primary indicator to evaluate a school as failing or proficient in their cognitive outcomes, we focus here on this test.

References

Andrews Paulsen, C., Ferrara, S., Birns, J., & Joffre Leclerc, K. (2002). *Multiple measures for student assessment and accountability in Massachusetts*. Concord, MA: American Institutes for Research.

Baker, E.L (2003). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, *22*(2), 13–17.

Barber, M. (2004). The virtue of accountability: System redesign, inspection, and incentives in the era of informed professionalism. *Journal of Education*, *185*, 7–38.

Booher-Jennings, J. (2005). Below the bubble: ''Educational triage'' and the Texas Accountability System. *American Educational Research Journal*, *42*, 231–268.

Chester, M.D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice*, *22*(*2*), 32–41.

Chester, M.D. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, *24*(4), 40–52.

Cullen, J.B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system* (NBER Working Paper No. 12286). Cambridge, MA: National Bureau of Economic Research.

De Wolf, I.F., & Janssens, F.J.G. (2007). Effects and side effects of inspections and accountability in education: An overview of empirical studies. *Oxford Review of Education*, *33*, 379–396.

Figlio, D.N., & Getzler, L.S. (2002). *Accountability, ability and disability: Gaming the system* (NBER Working Paper No. 9307). Cambridge, MA: National Bureau of Economic Research.

Gong, B., & Hill, R. (2001, March). *Some considerations* of *multiple measures in assessment and school accountability*. Paper presented at the Seminar on Using Multiple Measures and Indicators to Judge Schools' Adequate Yearly Progress Under Title I (sponsored by CCSSO & US DOE), Washington, DC.

Hamilton, L.S., & Koretz, D.M. (2002). Tests and their use in test-based accountability systems. In L.S. Hamilton, B.M. Stecher, & S.P. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 13–49). Santa Monica, CA: RAND. Retrieved from http://www.rand.org/content/dam/rand/pubs/monograph_reports/MR1554/MR1554.ch2.pdf

Hanushek, E.A., & Raymond, M.E. (2002, June). *Lessons about the design of state accountability systems*. Paper prepared for "Taking Account of Accountabiliy: Assessing Policy and Politics", Harvard University, Cambridge, MA. Retrieved from http://edpro.stanford. edu/hanushek/admin/pages/files/uploads/accountability.Harvard.publication%20version. pdf

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.

Jacob, B.A., & Levitt, S.D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *118*, 843–877.

Janssens, F.J.G., Vanotterdijk, R., & De Wolf, I.F. (2005, May-June). *De stand van het toezicht in Nederland en Vlaanderen* [Describing school inspections in Flanders and The Netherlands]. Paper presented at the Onderwijsresearchdagen, Gent, Belgium.

Koretz, D.M. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, *22*(2), 18–26.

Koretz, D.M., McCaffrey, D.F., & Hamilton, L.S. (2001). *Toward a framework for validating gains under high-stakes conditions* (CSE Technical Report 551). Retrieved from http://www.cse.ucla.edu/products/Reports/TR551.pdf

Ladd, H. (2007, November). *Holding schools accountable revisited*. 2007 Spencer Foundation Invited Lecture in Education Policy and Management at the Association for Public Policy Analysis and Management research conference, Washington, DC.

Lee, J., & Coladarci, T. (2002). *Using multiple measures to evaluate the performance of students and schools: Learning from the cases of Kentucky and Maine*. Orono, ME: University of Maine.

Stecher, B.M. (2002). Consequences of large-scale, high-stakes testing on school and classroom practices. Tests and their use in test-based accountability systems. In L.S. Hamilton, B.M. Stecher, & S.P. Klein, (Eds.), *Making sense of test-based accountability in education* (pp. 79–100). Santa Monica, CA: RAND. Retrieved from http://www.rand.org/content/dam/rand/pubs/monograph_reports/MR1554/MR1554.ch4.pdf