Paper Title Responses of New York City Elementary Schools to Multiple Measure Accountability Systems

Author(s) Melanie Ehren, Institute of Education; Thomas C. Hatch, Teachers College, Columbia University

Session Title High-Stakes Testing and Accountability: Responses and Effects

Session Type Roundtable Presentation

Presentation Date 4/27/2013

Presentation Location San Francisco, California

Descriptors Accountability, High Stakes Testing, No Child Left Behind

Methodology Qualitative

Unit Division L - Educational Policy and Politics

Responses of schools to accountability systems using multiple measures;
The case of New York City elementary schools[1]

M.C.M. Ehren
T. Hatch

ABSTRACT
Many studies point to potential unintended consequences of accountability systems such as when schools narrow their teaching to fixate on tested subjects. As a result, some states and districts in the U.S. have complemented the federal test-based accountability system with additional measures of educational practices to hold schools accountable on multiple measures. To explore the consequences of such systems, this study focuses on the responses of elementary schools to a multiple measure accountability system in New York City, including high stakes tests and quality reviews. While some schools showed broader improvement efforts, results suggest the state test remains the dominant measure in driving responses of schools, and, in some cases, the quality review further reinforces the schools' focus on the test.

INTRODUCTION
Numerous educational reform initiatives over the past dozen years have aimed to hold schools more accountable (Leithwood and Earl, 2000). In many of these initiatives, schools are held accountable for the quality and output of their performance and information on their performance is disclosed to policymakers and the wider public. Accountability systems often set targets on minimum performance levels and schools face consequences when not meeting these targets. As a result, schools are likely to feel substantial pressure to act in response to the accountability demands. These responses may be positive when teachers work harder or more effectively, providing more instructional time or covering more material, or when they redirect their resources to standards of teaching and learning that are seen as important. Such responses may lead to beneficial effects on learning and to valid increases in test scores. However, these responses can also be negative when they inflate test scores or harm student learning. For example, when teachers narrow their instruction to meet accountability measures, they may morph classroom practice into something measurable and auditable that produces good results on accountability measures, without leading to high quality instruction or reaching high standards on outcomes that are not measured (Koretz, McCaffrey and Hamilton, 2001; Stecher, 2002; Booher-Jennings, 2005).

These problematic and narrow responses 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. Several scholars therefore express 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 in this context refers to the use of two or more different methods to assess a number of different indicators of educational quality and performance of schools, rather than the use of a single measure several times (Baker, 2003; Gong and Hill, 2001). These multiple measures may include cognitive outcome measures (such as results of student achievement tests, aggregated to the school level), non-cognitive outcome measures (e.g. attendance and dropout rates) and other direct measures of educational practices (such as quality reviews or school inspections) (Koretz, 2003). These multiple measures are expected to motivate schools to focus on a broader set of

[1] The research reported here was supported by the Spencer Foundation

goals and to better connect to concerns such as student engagement in learning and instructional quality (Mintrop and Sunderman, 2009).

Despite this focus on the ways in which multiple measures may mitigate undue influence of single measures, multiple measures may have a variety of unintended consequences that complicate efforts to broaden school goals. Jos and Tompkins (2004) for example point out that multiple measure accountability systems may create a thick web of overlapping, path-dependent, layered relationships that often pose inconsistent and conflicting demands on schools. This condition may be created when newer decentralized accountability demands are added to centralized demands already in place. Conflicts may also be created when traditional accountability systems, particularly those that focus on ensuring compliance with rules and procedures, are complemented with newer measures of student performance. Conflicts also arise when schools face accountability demands from different authorities. The accountability requirements of the No Child Left Behind (NCLB) Act of 2001, for example, have been added on to many state systems, resulting in mixed messages regarding the performance of schools (Linn 2005). In some of these cases, such as in New York City where schools face both a state system based on NCLB and a city system that uses a school grading system, schools may fare well under one accountability system but poorly under the other (Pallas & Jennings, 2009). Under conditions where multiple, accountability measures are in conflict, or are unclear or ambiguous, schools may respond defensively or may seek out the most expedient or obviously acceptable position, preventing them from learning and trying out new solutions (Tetlock et al, 1989 and Ebrahim, 2005).

Adding to the uncertainty around the effects of multiple measures, rather than being used to prevent a narrow focus on tested subjects, multiple measures could also be used to strengthen the focus on specific outcomes and to align many aspects of the work of schools or educational systems around those outcomes. For example, in recent years, the New York City Department of Education has developed quality reviews and school surveys to complement the New York state test that is used to measure student performance and student progress in schools. These additional measures intend to focus responses of schools and to align school improvement efforts, curricula, assessment and instruction to the tested standards. The school survey is designed to help school principals understand how to improve the school's learning environment with the purpose of accelerating academic achievement. Schools are expected to use the survey results to identify their strengths and weaknesses, to develop goals and interim benchmarks for progress and to identify strategies to achieve these goals[2]. The quality review process was also implemented to communicate and reinforce a set of behaviors and practices that are expected to drive improvements in student achievement. The reviews are expected to provide schools with different, timelier data that could serve as a *leading* indicator of how their students might perform in the future so that they could adjust their instructional approaches and management systems in order to continuously improve.

In short, although multiple measures in accountability systems may be seen as a means of mitigating the narrowing effects of single measures, the possibility that multiple measures may have other effects remains to be more fully explored. This study aims to fill this gap by studying responses of New York City elementary schools to the multiple accountability measures in place in 2010-11. These measures have changed substantially over the past ten years, but in 2010-11, these measures included school quality reviews of educational practices and a school progress report that focused primarily on students' test scores but took into account results from surveys of

[2] Survey results to improve schools; worksheet for school leaders; School survey code of ethics; Educator Guide to NY City Progress Report

teachers and parents as well. In order to explore the effects of these multiple measures on New York City schools, this study focused on two questions:

- How do schools respond to multiple measures in an accountability system?
- How do these measures interact to promote broad/narrow responses of schools?

Answers to these questions will illuminate the conditions under which multiple measures in accountability systems may better fulfill the goals to improve policy, school leadership and classroom instruction.

THEORETICAL FRAMEWORK

This section first outlines responses of schools to different types of accountability measures. Available studies on responses of schools to educational accountability so far primarily include single measure systems, where school improvement and changes in student achievement are measured resulting from either test-based accountability systems or school inspections/quality reviews. Although these results do not show how schools will respond to multiple measure systems, they are indicative of the range of responses multiple measure systems may produce.

Responses of schools to test-based accountability

Evidence of current responses of schools to accountability measures of cognitive outcomes can particularly be found in the U.S. where test-based accountability is the dominant form of accountability in education. Several studies provide descriptive taxonomies, describing effective, ambiguous and ineffective responses by teachers and principals to test data and illuminating the potential consequences of these responses (Mehrens and Kaminski, 1989; Popham, 1991; Haladyna, Nolen, and Haas, 1991; Koretz, McCaffrey and Hamilton, 2001; Stecher, 2002; Volante, 2004). Ambiguous and ineffective responses fit our description of narrow responses.

Effective responses include working harder or more effectively to cover more material. Responses have a more ambiguous character when school staff 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. Teachers may shift their instructional time between subjects (for example from science and social studies to language and math) or between topics within a subject. Examples include pushing instruction towards lower order cognitive measured skills if standards do not also emphasize higher-order skills or neglecting other content not explicitly addressed in the standards.

Ambiguous responses also include teachers coaching their students to do better by focusing instruction on narrow, specific aspects of the test that are partly or entirely incidental to the definition of the domain the test is intended to represent (Stecher, 2002). They spend time familiarizing students with item formats or scoring rubrics, or use old tests in instruction to teach test-taking strategies. Koretz (2003) describes an example of teachers noticing that the high-stakes test always uses regular quadrilaterals and triangles in area and perimeter problems; correspondingly, those teachers may decide not to use irregular polygons or figures with more than four sides in instruction. In order to improve a school's overall score on the accountability measures, teachers sometimes also target instruction to students close to a cut-point (educational triage) (Booher-Jennings, 2005; Gribben et al, 2008). For example, teachers might use test scores to divide students into three groups: safe cases, suitable cases for treatment and hopeless cases. They may then target resources to those students most likely to improve the school's scores—the "suitable cases for treatment" or "bubble kids". According to Booher-Jennings (2005) low scoring students are given the least attention during the course of the school year and receive increasingly less time to become proficient as the year goes on. Students at the far ends of the

academic spectrum are also neglected. Resources are distributed to increase aggregate test scores rather than meeting the needs of individual students.
Ineffective responses occur when schools and teachers distort data and cheat to improve the school's status on the accountability measures. Figlio and Getzler (2002) and Cullen and Reback (2006), for example, describe how some schools at risk of failing improve their state-assigned grade or classification by taking their poorest performing students out of the testing pool and "reshaping the test pool". Schools may do so by classifying (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 exemptions from testing by parents of low achieving students and increasing dropout rates of low achieving students. Jacob and Levitt (2003) found that 4-5% of tested classrooms show evidence of some form of cheating on standardized tests each year. Cheating may involve teachers prompting students with the right answer during a test, providing the actual test items in advance, providing hints during test administration, making changes to answer sheets before scoring or leaving pertinent materials in view during the testing session. Teachers have opportunities to cheat on the tests when they are responsible for administering and scoring the test. The test may in itself be free from measurement bias but the method of administering includes potential biases.

Responses of schools to school inspections
Responses of schools to school inspections and quality reviews have primarily been studied in Europe where inspections of schools are the dominant accountability measure. Schools respond to school inspections when preparing for visits or when acting on the feedback provided during or as a result of these visits. A recent literature review by Klerks (in prep.) summarizes the effects of school inspections on behavioral change of teachers, school improvement and student achievement results. Her study of empirical research (with a high score on the Maryland Scientific Methods Scale, including quasi-experimental research designs) and published in peer reviewed journals since 2000 shows plausible connections between school improvement and behavioral change of teachers, and small (both positive and negative) causal effects of school inspections on student achievement results.

Klerks' literature review shows that no specific characteristic of school inspections in itself leads to improvement, but effects arise from a complex interaction between inspection characteristics and students, teachers and the school management. Gray (cited in Visscher and Coe, 2002, p. 2), Kogan and Maden (1999) and Ehren and Visscher (2008) for example describe how schools use school inspections, and the feedback provided during inspections, to implement improvements such as rules of conduct for students, strategies for raising examination results, changes in monitoring and assessment of students and changes in management styles and structures. Gray (cited in Visscher and Coe, 2002, p. 62), based on an in-depth study of twelve schools, found three different routes of school improvement after external evaluations: tactical improvement aimed at improving student performance, strategic thinking aimed at developing school policies and classroom activities and finally capacity building, which is only carried out by a few schools. These schools improve continuously by pulling all relevant levers for change, including learning from classroom experience and encouraging staff professional development.

A number of studies however also point to unintended consequences of school inspections. Rosenthal (2004) for example found a decrease in examination results of pupils in England in secondary education in the year of the inspection visit. These negative consequences may result from intended and unintended strategic behavior of schools and teachers, which have been categorized by De Wolf and Janssens (2005). They describe intended strategic behavior as schools trying to manipulate the inspection assessment through window dressing,

misrepresentation or gaming. Window dressing refers to schools implementing procedures and protocols that have no effect on primary processes in the school, but are implemented to be assessed more positively. Schools 'brush up' to receive a more positive assessment. Ehren (2006) for example found that some Dutch schools counted time spent outside playing as part of lesson schedules to comply with the minimum number of lesson hours. Misrepresentation also occurs when schools manipulate behavior on which they have to provide reports. Examples are excluding low performing students from exams that are used to assess schools. Gaming refers to schools manipulating actual behavior instead of reported behavior. For example, Schools may attempt to lower the targets by trying to perform low in the year the targets are set. Another example of gaming was found by Chapman (2001) who showed that teachers prepared and structured their lessons better when inspectors visited the school. During inspection visits, teachers also taught in a more structured, classical way and refrained from having pupils work together in small groups, which could cause disruptions.

Unintended strategic behavior refers to the influencing of behavior by the assessor and/or by the method of working used for the assessment. In effect this means a one-sided emphasis on the elements that are assessed. Schools for example may emphasize phenomena that are quantified in the performance measurement scheme, at the expense of unquantifiable aspects of performance or risky innovations. Schools aim at success that can be established very quickly, instead of long-term school improvement, or they focus on implementing measures of success (e.g. procedures and protocols for self-evaluations and data use) rather than the underlying objective (improving student learning). Thus, school inspections may lead to narrow responses when schools act strategically when trying to manipulate the inspection assessment or one-sidedly emphasize the aspects of their education that are being assessed.

Narrowing/broadening of responses in multiple measure systems

This short literature review of schools' responses to single measure test-based accountability and school inspections shows that each approach, on its own, can lead to ineffective responses. In particular, either approach may lead to narrow responses. Responses to testing are considered to be narrow when instruction focuses on the specific content and format of the tested material or specific subgroups of students for which accountability targets are set, rather than the full domain of knowledge and skills represented in state standards or all students in a school. Such responses have the potential to inflate test scores and to harm student learning. School inspections may lead to narrow responses when schools try to manipulate the inspection assessment or one-sidedly emphasize the processes and practices that are the focus of the inspection.

The use of multiple measures, however, may help to mitigate the narrowing effects of single measures. In particular, including measures of both cognitive outcomes (tests) and educational practices (school inspections/quality reviews) in a multiple measure accountability system may broaden responses of schools in a number of ways. First of all, adding measures of educational practices (e.g. school inspections) to test-based accountability systems can prevent schools from fixating on a small number of indicators as represented by the test. Second, adding the use of test results as an output measure of schools to school inspections can discourage schools from simply implementing the processes and practices that are the focus of the inspections without taking into account data on the extent to which those practices may be improving student learning.

In this study we therefore define broad responses as setting goals, using curricula and making instructional decisions that reflect variables and outcomes that include but are not limited to those that are directly assessed on a single measure (such as the test or the inspection framework). For example, schools may set goals for improvement in a wide range of subjects, instead of only tested subjects. Similarly, they may equally allocate instructional time and

resources to a wide range of subjects (instead of only tested subjects), and they may try to make improvements in a wide range of educational practices that are not directly related to improving test scores in the immediate future (e.g. cooperation between teachers, quality assurance, safe learning climate in the school), but could lead to improved student achievement in the long run.

Several authors refer to mechanisms of institutionalization or refer to the social psychology of accountability to hypothesize on the likeliness of multiple measures leading to such broad responses. In particular, Honig and Hatch (2004) explain how schools can engage in a process of "crafting coherence" to use external demands such as those from accountability systems to advance school goals that go beyond those highlighted in the accountability instruments themselves. Crafting coherence is a process of negotiation in which school members use bridging and buffering activities to take advantage of demands that support school goals and to ignore or adapt those that may constrain school goals and practices. In order engage in this process, schools draw on their knowledge of the indicators and instruments of accountability, of those to whom they are accountable and their own beliefs and current practices to decide when and how to conform to external demands and to shape their responses (Ebrahim, 2005; Booher-Jennings, 2005). Multiple measures can support this process by giving schools an opportunity to connect to and build on a wider range of indicators.

In this study, we focus particularly on the outcome of such organizational decision-making processes, that is: *the actual changes schools (principals and teachers) make in their goals, curriculum and instruction and whether these changes reflect responses to the broad set of multiple accountability measures in New York City or whether schools single out changes to specific measures.*

Below we will first describe the multiple measures that are in place in New York City.

Accountability measures for New York City Elementary Schools

The New York City Department of Education (DOE) is the largest system of public schools in the United States, serving about 1.1 million students in nearly 1,400 schools, with approximately 135,000 employees and an annual budget of $23 billion. Over the past ten years, the DOE has made numerous changes in their approach to accountability and (K-5th grade) elementary schools in New York City now face a complex array of accountability measures and demands. These include measures based on the Federal No Child Behind Legislation that New York state uses to hold schools accountable for their performance and a set of measures that the New York City Department of Education has designed as part of their own accountability process. The following section provides a description of the accountability measures, targets and consequences during the time of our study in 2010-11.

Measures of cognitive outcomes

In New York City, cognitive outcomes of elementary schools are measured annually by means of student achievement tests in grade 3 to 5. Students are tested on the New York State standardized test to assess their mastery of the New York State learning standards in ELA, math and science (science is only tested in grade 4). The results of the test are used to assign students to one of four performance levels, reflecting the extent to which the student demonstrates the level of understanding expected at his or her grade level (level 3 is “proficient” and level 4 “advanced proficient”). The performance levels in each subject are indicated by a test score range; students are for example proficient in science if they score between 65 and 84.

Students’ scores on the state test are used by New York State to implement federal NCLB requirements and to determine if schools meet the state goal of having every student at or above

the proficiency level in reading and mathematics by 2014. To make sure schools reach those goals, New York state and NCLB require schools to meet an Annual Yearly Progress (AYP) target every year. Schools make Adequate Yearly Progress if they meet the participation and performance criteria in English Language Arts, mathematics in grade 3-5, and Science (grade 4). Adequate Yearly Progress is determined separately for each accountability subgroup of Asian, Black, Hispanic, Native American, White, multiracial, low-income, limited English proficient and special education students when these subgroups include 15 or more tested students. All these subgroups must make adequate yearly progress in math and ELA (and science grade 4) for the school to make adequate yearly progress in that subject.

The number of years a school fails to make adequate yearly progress and the number of subjects and accountability groups for which a school fails to make adequate yearly progress are used to assign a school to an accountability phase and category. The greater the number of accountability groups that failed to make AYP, the more intense the intervention. Interventions include an educational assessment by a joint intervention team, with schools expected to use the findings of the review, audit or assessment to develop an improvement plan. Schools not making adequate yearly progress for two consecutive years have to provide students with supplemental education services (such as free tutoring), and take corrective actions in addition to providing students with other school options.

In addition to the NCLB-based measures New York state uses to hold schools accountable, the New York City Department of Education uses its own system of progress reports for holing schools accountable. The progress report uses a complex metric to give each school a letter grade (A through F). The letter grades are based on a formula that includes the weighted value of *student performance (*25% of the overall score), *student progress* (60% of the overall score) and *school environment* (15% of the overall score). Student performance and student progress are based on students' scores on the New York state test. Schools can also earn additional credit when their high-need students achieve exemplary outcomes. The school's environment score is based on attendance rates and the results of a parent and teacher survey (see section on measures of educational practices).

In general, if no progress is made over time and schools receive an overall grade of D or F will be subject to school improvement measures and target setting and possible leadership change (subject to contractual obligations), restructuring, or closure. The same is true for schools receiving a C for three years in a row. Ultimately, schools are accountable for making progress and receiving an overall grade of A, B, or C. The Progress Report Grade is also used as part of the principal's evaluation, counting as 32% of the annual formal evaluations.

Measures of non-cognitive student outcomes

New York State and New York City both collect school records to measure non cognitive outcomes of schools. These records include data on student attendance on each school day in elementary schools and participation of subgroups of students in ELA, math and science on the state test. New York State also requires schools to provide data on student suspension for publication in the school report card. Considered part of the school environment measure, attendance rates count as 5% of the New York City Progress Report grade.

Measures of educational practices

Both New York State and New York City also measure educational practices in schools. Schools have to provide data on their average class size, qualifications of their teachers, teacher turnover rate and staff counts to New York State. This information is published in the New York State school report card.

Elementary schools in New York City are also held accountable for their educational practices through the use of a parent and teacher survey. The survey collects information on the school environment and the results are used for 10% of the school's overall score on the Progress Report grade. The survey is administered annually to parents and teachers and gathers information on how well the school creates an environment conducive to student learning. The survey includes the following four categories: 1. high academic expectations of students and providing a learning environment to promote academic success, 2. communication of educational goals and requirements within the school and to the school community, 3. engagement of students, parents and educators in student learning (through curricular activities etc.), 4. a safe and respectful learning environment. A school's results are compared to results of all schools serving the same grades throughout the City and to a peer group of schools with similar demographics[3]. A school's overall score is then assigned a percentile ranking based on the range of all scores Citywide, by school type.

In addition, the New York City Department of Education evaluates educational practices of elementary schools through a teacher and a parent survey (which count as part of the school environment score on the NYDOE progress report) and through school quality reviews (which do not count as part of the progress report grade). The focus of the Quality Review is to assess the quality and performance of a school in relation to New York City Department of Education's Quality Criteria. Reviewers evaluate the key aspects of the school's work to determine how well they align with each other and with the five areas of the Quality Review rubric – all in service of improved student outcomes[4].

Schools are generally scheduled for quality reviews once every three to four years, but poor performance on the progress report or on the State's accountability measures may also lead to a quality review. The focus of the school quality review is a school visit which takes place on two- or three-days and is carried out by one reviewer who is appointed by the New York City Department of Education. Reviewers are drawn from a pool of educators including community and high school superintendents, Quality Review Directors, and other administrators or instructional leaders associated with the Department of Education. Before the visit, the school produces a self-evaluation in advance of the review. The self-evaluation is based on a standardized form and a quality review rubric. The rubric contains five quality statements on 1) instructional and organizational coherence, 2) gather and analyze data, 3) plan and set goals, 4) align capacity building, 5) monitor and revise. These quality statements are detailed in indicators. Each indicator is evaluated on a four point scale: well developed, proficient, underdeveloped with proficient features, and underdeveloped. Reviewers draw upon the school self-evaluation and school data during the school visits. Based on the conversations they have with principals, teachers, students, and parents during the school visit and their visits to classrooms and observations the reviewers uses the rubric to evaluate how well the school is organized to educate its students.

Quality reviews are scored using a numeric system earning points for each indicator that sum up to a final score. The score of at least three of the indicators are used to score a quality statement. The overall score is based on the overall assessment of the quality statements; e.g. a school is scored proficient when at least four quality statements are proficient, and all quality statements are at least underdeveloped with proficient features. A score range of 92-100 represents 'well developed', 72-91 is proficient, 47-71 is developing and 25-46 is underdeveloped.

[3] Three-quarters of a school's score in a given area is based on a school's position relative to other peer schools, whereas one-quarter of a school's score is based on how well a school does relative to other schools across New York City. Source: Accountability reforms before and after mayoral control

[4] Quality reviewers handbook; http://schools.nyc.gov/Accountability/tools/review/default.htm

After the site visit, schools receive a quality review score and a report is published on the website of the New York City Department of Education. This public information is designed to be used by parents and others to influence their views of the school. The quality review score (well-developed, proficient, developing, under developed) represents a separate accountability score that is included in the Progress Report but it is not used as part of the grade calculation. Schools that score 'underdeveloped with proficient features' or 'underdeveloped', and have a progress report grade D or F, are subject to possible leadership change or school closure (Pallas and Jennings, 2009). The Principal Performance Review will incorporate the school's most recent Quality Review score. The Quality Review counts for 22% of the total principal's evaluation[5].

The table below provides a summary of the accountability measures of New York City elementary schools.

[5] Principals guide to the quality review

Table 1. Summary of accountability measures for New York City elementary schools

New York State accountability			New York City accountability		
Performance area	Type of measures and sources of data	Target and consequences	Performance area	Type of measures and sources of data	Target and consequences
Cognitive outcomes: Student achievement in grade 3-5 in ELA, math and science	Annual standardized student achievement tests, using m.c., short and extended response questions	Adequate Yearly Progress: - Participation criteria: valid scores of 95% of students in ELA or math, and 80% of students in science in grades 3-6. - Performance index in ELA and math of subgroups of ≥ 30 students must be equal or greater than the annual measurable objective, or schools must make safe harbor. - Performance index in science (grade 4) of subgroups of ≥ 30 students must be equal or greater than the state standard (100) or meet the progress target. Schools that fail to make AYP for a subgroup of students in a subject face interventions. The greater the number of subgroups and subjects for which the school does not make AYP, the more intense the interventions.	Cognitive outcomes: Student achievement in grade 3-5 in ELA, math and science	Annual standardized student achievement tests, using m.c., short and extended response questions	Progress report grade: A, B is scoring above the threshold; scoring C three years in a row, D or F is scoring below the threshold. The grade is comprised of: - School environment (15%), including survey score on school environment (10%) and attendance rates (5%) - Student performance (25%), including (weighted) percentages of students at proficiency in ELA and math, and median student proficiency in ELA and math - Student progress (60%), including

		The school moves to a next phase of increased interventions after two years of not making AYP.			(weighted) median growth percentages in ELA and math, and median growth percentile for school's lowest third students in ELA and math The Progress Report Grade counts to 32% of the annual formal evaluation of principals. The quality review score counts to 22% of the formal annual evaluation of principals. A Progress Report Grade of D and F will eventually lead to restructuring or closure of schools or leadership change.
Non cognitive outcomes: attendance and suspension rate, participation rate on state test	Records of schools		Non cognitive outcomes: attendance and suspension rate, participation rate on state test	Records of schools	
Educational practices: indicators on teacher qualifications, teacher turnover rate, staff count in report card	Records of schools		Educational practices: - quality statements in review rubric (instructional and organizational coherence, gather and analyze data, plan and set goals, align capacity building, monitor and revise)	- Quality reviews - Parent, teacher and student survey	

METHODS

We used nine case studies to explore how schools respond to the accountability measures used by both the state and the city. We selected schools with different grades on the state and City accountability measures as their prior performance on these measures will likely affect their current responses to the measures. Two to three schools were selected for each school grade A, B, C or D based on their progress report in the previous year (2009-10). Most of these schools made the state-wide AYP target in 2009-2010, but two schools failed to do so for one or more years. Illustrating the conflicting nature of the state and city accountability systems, one of the schools that did not make AYP received a B on the city's progress report and the other school received a C. The two schools in the sample that received a D on their progress report made AYP. In general, differences in these two accountability measures reflect the major emphasis in the city progress reports on growth in test scores.

Within each school, the principal and a teacher in grade 2 and in grade 4 were selected for interviews. These teachers were expected to respond differently to the accountability measures as the state test was administered in grade 4 and not in grade 2.

All of the case study schools received a quality review during the period of our study; no schools with an F on their Progress Report were scheduled for a quality review during our study and therefore none were included in the cases. Quality reviewers of schools we labeled A and C were reviewed by a leader from the network that provided them with support.[6] The other seven schools were reviewed by their superintendent. The distinction is relevant as superintendents have additional tasks in the schools they review, such as evaluating performance of principals, approving the school's comprehensive education plans (outlining improvement goals for the upcoming year), deciding on tenure of teachers and on promotion of students in testing grades.

Table 2 provides an overview of selected schools.

Table 2. Case study schools

		New York City target			
		Progress report Grade A	Progress report Grade B	Progress report Grade C	Progress report Grade D
New York State target	Made AYP for all student groups in all subjects: in good standing	School A School B	School C School D	School F	School H School I
	Failed to make AYP in one/more student groups in one/more subjects		School E	School G	

Table 1 in the appendix provides a summary of background information on the schools.

We collected data on responses of these schools to the accountability measures by means of a document analysis, interviews with principals and two teachers and observations of the quality reviews. Documents reviewed included the quality review reports, the school's Comprehensive Education Plan (describing the school's goals for the upcoming year), pacing calendars and the school's self-evaluation report. The documentation of the quality review included observations of approximately 6-10 lessons that were observed by the quality reviewer; observations of several

[6] All schools in New York City are expected to join a network that provides a range of services to schools, particularly support for instruction and professional development.

meetings of the quality reviewer with the principal and assistant principals, with individual teachers and teacher teams; and observation of the feedback forum at the end of the two day visit in which the quality reviewer explained the quality review score to the administrative team and the network leader of the school (who supports the school in professional development and school improvement). The following table provides an overview of the data collection.

Table 3. Overview of data collection

		Method of data collection		
		Document analysis	Observation of quality review	Interviews with principal and two teachers after review
	School's performance on measures of cognitive outcomes (school's performance on accountability target and consequences)	*		*
	School's performance on, and description of measures of educational practices (school's performance on QR rubric and consequences, feedback, relationship and communication reviewer-school)	*	*	*
	Responses of schools to cognitive outcome measures (goals, curriculum, instruction)	*	*	*
	Responses of schools to educational practices measures (goals, curriculum, instruction)	*	*	*

The interviews were transcribed and data was analyzed using the program Atlas/ti (using the variables as codes to analyze the transcripts). For each school a case study report was made in which a description was given of the school's performance on each of the accountability measures, the feedback they had received in the quality review, relevant internal and external conditions in the school and the school's and teachers' responses to the accountability measures. In the interviews we specifically asked principals and teachers how the state test and the quality review informed school goals and policy, curriculum and instruction. The documents and observations of the quality review were also analyzed to identify such responses. Only responses that were indicated in two or more sources were included in our results section to enhance the reliability of our results and limit potential bias due to self reports.

Responses of teachers and principals to both the state test and the quality review were identified as broadening when principals or teachers described goals, school policy, or (decisions about) changes in curriculum and instruction that were not directly related to tested subjects or test-related targets, such as implementing assessments of students' knowledge and skills in social studies. Responses were marked as 'narrow' when principals and teachers aligned their goals, school policy, curriculum and instruction to (only) tested subjects, subgroups, topics and item formats. The feedback that was provided during the quality review was also categorized as broadening/narrowing to explain the potential of quality reviews to broaden or narrow responses of schools. Feedback was considered to potentially broaden responses of schools when the quality reviewer addresses necessary changes that go beyond what the high stakes test requires.

A description of all the variables in each school was included in a case study report. The case study reports were used to carry out a within case analysis in which the relation between the variables in our study were analyzed to describe how each school and teachers within the school responded to the accountability measures and how internal and external conditions affect these responses.

RESULTS

Below we first include a summary of the goals, the curriculum and the instruction in our case study schools and which measures inform such decisions. Next, we will separately describe how schools responded to the quality review and whether these responses broaden or narrow their goals, curriculum and instruction. Finally we describe the aspects of the quality review (rubric, feedback, role of the reviewer) that motivate such broad or narrow responses.

Goals

We analyzed schools' comprehensive education plans (CEP) and asked principals and teachers about the goals of the school, and how the test and the quality review informs these goals. The results show that all schools have goals in their comprehensive education plans on improvement of student achievement, as measured in the state test and in other interim benchmark assessments. Principals explain that superintendents require schools to at least include goals on math and ELA in their comprehensive education plans. One of the principals for example states that the CEP:

> 'definitely needs to have the academic goals there. Because you are told that they need to be there. So you cannot create a comprehensive education plan without having a goal for ELA and Math so those two always have to be there. And obviously my goal will always going to be to do better than we did last year.' (school E)

The goals include percentages on improvement of student performance in ELA and Math which are based on the school's prior performance on the state test, or on other (formative) assessments. They often (in all schools, expect school A) reflect the improvements in test scores of (subgroups of) students to enable the school to make AYP and/or increase their progress report grade. Principals try to set realistic percentages in their goals, as these goals are used in their annual performance reviews.

School I for example aims to increase the performance in math by reducing the percentage of level 1 students by 5%, and increasing the percentage of level 3 students by 5%. School E aims for an increase of 3% of students with disabilities and limited English proficient students performing at level 3 or 4 on the state test.

Some schools also have goals on improvement of student achievement in social studies and science (school D and H), school improvement in other areas (e.g. the use of technology in teaching, professional development of teachers,parental involvement, student attendance) (school A, C,D, E, F, H) and improvement of all students (instead of only targeted students) in math and ELA (school C, F, H). These goals reflect the schools' mission on for example providing a broad curriculum for all students (school A), or schools feel improvement of student achievement of all students is needed to make AYP (school C, H).

The quality review informs schools' goals to some extent through the feedback and line of questioning of quality reviewers. In the majority of the schools (school E, I, H, D and A), the quality reviewer specifically questions the school or provides feedback to the school about goals to improve student achievement in math and ELA. Quality reviewers for example highlight the subgroups of student schools should target to receive additional credit on the progress report or meet AYP, they question the schools on what goals they have in place to improve student achievement in specific areas or for specific student groups (e.g. Black students, or IEP students) or motivate the school to include specific numbers and percentages by grade and subgroup on improved test scores. In one school (school H), the quality reviewer motivates the school to only focus on ELA and math goals and delete goals on improvement of student achievement in other subject areas from the school's comprehensive education plan.

Curriculum

In the interviews we asked principals and teachers how they decide on the choice of units in their curriculum, the length of each unit, when the unit is scheduled during the year and how the test and quality review inform these decisions. Principals and teachers mention different sources of information and guidelines that inform their curriculum, such as the State standards, textbooks, or materials from the Teachers College Readers and Writers Project, their network or from the internet. All the schools work from an existing curriculum that they have selected that has been in place for several years. At the end of each academic year (May/June) they evaluate the curriculum, often making minor adjustments and establishing the schedule for the next academic year. When explaining how they decide on these revisions, school members frequently talk about how the test (the topics in the test and students' performance on the test) informs their decisions on changing the schedule of units of study, extending units of study (supplementing the curriculum) or changing the topics that are covered within units of study.

All schools (except school C and G) have analyzed the state test to learn which units of study they need to cover before the test and have rescheduled units that will not be tested to be taught after testing; schools for example move the unit of study on 'bar and line graphs' to early on in the year as this unit is tested. Schools also align units of study in ELA to specific genres (e.g. non-fiction, informational texts) that are on the test to familiarize students with the genres they will be reading on the test and will be answering questions about.

According to principals and teachers in school C, D, I and H, units of study that get the most items on the test (e.g. non-fiction and informational texts) get more teaching time on the schedule, while schools D and C add instructional material to the curriculum that they are using to address tested skills that are not specifically covered in the textbooks. Schools using the EDM textbook[i], which uses a spiraling curriculum where one unit of study covers multiple skills as represented on the test, for example often supplement this curriculum with material on 'place value' as this topic is covered on the test but is not part of the curriculum. A teacher in grade 2 for example explains:

> Okay well normally we start planning at the end of May, June, and we look at how much time we spend on each unit, so for example, we'll say, the children have read notes, children have a difficult time, let's say, non-fiction texts, so maybe we'll cut down, maybe 2 weeks, with something that is easy for them like fairy tales, and we'll pull out more for non-fiction, because right now everything *(on the test)* is really non-fiction, non-fiction. (school I).

All schools, except school B, G and I also use students' test scores in the current year, or on an old state or predictive test at the start of the academic year to decide on which topics to teach or to emphasize in the upcoming year. As a grade 2 teacher explains:

> We looked at the third grade test in math and we noticed that a lot of the students weren't performing well in place value.. And then we looked at all the classes and we saw that in everyday math, the math program we use, there really hasn't been a lot of explicitly taught place value lessons. And then we looked at the common core standards and then we saw what the big emphasis place value is in this and that our math curriculum really wasn't addressing it. So as a second and third grade team we got together, we made all these explicit place value lessons, we got resources online to help the kids. So that's an example of seeing the test scores in the third grade and start targeting them earlier on. (school D).

Changes in the test and in the number of items for specific substandards, are picked up by schools very quickly and immediately lead to adaptations in the curriculum and teaching, often even before the new test has been administered. Schools frequently check online resources to inform themselves of scheduled changes to the test, the support network provides them with updates of changes in the test (referring them for example to sample tasks and item banks published by

companies such as Appleseed analytics), and available sample items are quickly distributed to teachers. A principal for example explains:

> We have changed our pacing calendars. Right now I am changing the benchmark calendar because in looking at the benchmark calendar the publishing dates for how long it is going to take kids to complete writing projects and we've changed when particular units are taught to make sure that the nonfiction units are taught before the test because we know that there is going to be more nonfiction on the tests. So it has affected the teaching, it has affected our benchmark calendar. (..) Our benchmark calendar is when key things are expected to be done and completed throughout the year. (School F).

Principals and teachers in all schools describe how ELA and math are the most dominant subjects in the curriculum in their school. Not only is most of the time on the schedule devoted to math and ELA, but in most schools, science and social studies are also somewhat integrated with instruction in reading and writing to increase teaching time in ELA. Schools F, D and G for example explain how they have students write about a social studies or science topic or read in these content areas to improve students' reading and writing skills. Schools feel this is good practice as the new common core standards emphasize (nonfiction) reading and writing in the content areas. A grade 2 teacher for example outlines:

> I push, reading and math in a different way like science and social studies, I try to incorporate the time of reading, by reading something that's related to science or something that's related to social studies.(school G)

Schools A and B emphasize the implementation of a broader curriculum, also including for example Arts and music. These schools express the need to provide students with a broad curriculum, and explain that their high performance on the state test and their highly educated parents who don't panic when the school's scores drop, allows them to offer such a broad curriculum. The principal in school A for example states

> Why should we ruin our curriculum with test prep so that this kid only gets two wrong instead of four wrong? They know the same amount of stuff and let's do more art, music and socials studies you know, and dance and not worry about that. And I think that that is a real disservice and I think that in a school like ours, we have the luxury of still being able to do that. Partly because our scores are high and secondly because our parents are savvy and understanding and are not going to flip if we get a B next year. (school A).

The quality review also informs schools' curricula, particularly through the rubric, the line of questioning and feedback on how schools use the use of data to inform the curriculum and how they align the curriculum to the state standards (particularly in math and ELA). Quality reviewers for example question the school about the curriculum they have in place for math and ELA and how the curriculum is adjusted to meet the needs of students (school I, C, H, D, A); they provide feedback on how student achievement data should be used to inform decisions about instructional content, emphasizing that schools should use benchmark assessments to monitor students' progress and adapt the curriculum. Quality reviewers in schools C and D emphasize that the curriculum should include more challenging academic tasks or tasks that address higher order thinking skills for specific subgroups of students and in school H that all students should be invited for Saturday Academy and Saturday morning test prep in ELA and math to further students' progress (referring to specific test programs/books that should be used). The quality reviewers in school I compliment the school for having a broad curriculum in place that addresses both the social-emotional and academic well-being of students, while the quality reviewer in school C motivates the school to accelerate the curriculum in kindergarten and monitor the social studies curriculum.

Instruction

We asked teachers to explain how they decide on what topics to teach within a unit of study, how to group students and how the test or quality review informs their instruction. Teachers and principals explain how these decisions are informed by the content, the format and students' results of/on the state test and on predictive benchmark assessments.

In New York, the Department of Education provides teachers with predictive benchmark assessments and item-skills analysis, displaying results from predictive benchmark assessments (which are aligned to the state test) for individual students and subgroups on specific skills within each standard. The item skills analyses provide detailed information of how much progress students need to make to reach (a higher) proficiency (level). Teachers are expected to discuss these results in inquiry teams and use the data to target specific skills that need to be retaught to specific (subgroups of) students.

Principals and teachers in our case studies explain how they use this information to target specific students for re-teaching that will support the school in making the state's AYP target. They use this information to decide on topics for instruction, grouping of students and how to explain instructional content and formative assessments. Additionally they talk about specific test preparation activities that are implemented before the state test. Each of these instructional activities is discussed in more detail below.

Choice of topics for instruction

Teachers in school C, E, F, H explain how they analyze the test and use an item skills analysis to decide on the topics to include in their instruction and the instructional priorities within each unit of study. The item skills analysis provides an overview of students' deficiencies in specific skills and strategies (e.g. 'adding and subtracting 3 digit numbers with and without regrouping), and the progress students are making (both individually, as by subgroups, class and grade) on periodic assessments, such as ACUITY, that predict students' performance on the state test. For example, a New York principal explains,

> We also have Acuity[1] on the periodic assessment that's provided by the DOE. And that's, you know, it's similar to the state exams. …It gives you data on how the children are performing and it lets you know what areas you have to focus your instructional attention on. The school wide assessments that we administer are the previous New York State exams. So to me that's really a good measure of what needs to be worked on. ... Mostly, we look at the data, then we break it down to the classrooms or the grades and that's the item analysis, we analyze the questions we have to reteach or we have to change our strategy, based on the questions and the outcome and the analysis. (school C)

Similarly, another New York principal notes,

> We created a spreadsheet where all the questions are laid out by skill, so say for example 'main idea' was question 2, 4, 6 and 20. So see how they did on those questions, so we thought that there were, let's say, they did not get the majority of the questions, they got even one wrong, it's already a reteach, it's already a reteach, if we saw that they have a whole topic or selection that they got correct, in some classes whole sections are correct, then that means that there's, we don't need to reteach. (school E)

A grade 2 teacher explains how she analyzed the state test in the subsequent grade 3 to inform her instruction:

> I noticed that some of the math questions were more complex, they were multistep, So at this point, after the February vacation, I try to include multi step math problems, as my problem of the day which was something that we do.

Grouping of students

Teachers in all schools also explain how they use the items skills analysis to homogeneously group students on three levels, reflecting their mastery/deficiencies in the same skills and strategies. Teachers may work with what they refer to as a 'lower functioning group', including students with proficiency levels 1 and lower 2, a middle level group including students who score a high level 2 and a low level 3, and high performing students who score a high level 3 and level 4. Each group is provided with differentiated small group instruction and different tasks. Teachers in schools A, B, C and H also use other criteria to group students, such as interest and learning profiles to group students. A grade 4 teacher explains:

> Well when we look at the… for example, I'm just going to refer to acuity or ARIS, when we go and look at it and we analyze the data we see specific skills based on the standards that our children missed or were getting correct, so we can make a grouping based on that skill and then that skill will become a small-group instruction. Sometimes it's a skill that most of the children have, and then it's a whole class lesson, or there's four, five children that missed that skill then we make it into small-group instruction and it's supported through small-group instruction (school H).

Explanation of instructional content

Teachers used specific aspects of the test (including rubrics, formats, and content) in instruction, with the goal of helping students gain familiarity with what the test looks like and how it is scored, and to learn about misconceptions of their own students when answering items on the state test. Teachers analyzed which distractors prior years' students had chosen when answering multiple choice questions and discussed these wrong distractors or one or more specific problems with their current students; teachers discussed specific items and explained rationales for particular responses; and teachers explained scoring rubrics and had students practice in scoring their own answers and those of other students, using the same rubrics as are used to score the state test. In addition, teachers also had students practice old test items during instruction.

Teachers for example focus on teaching students to express their thinking in writing instead of in writing conference (style, punctuation, capitalization) when the scoring rubric assigns more credit to these expressions then to writing conference. In all schools we found examples of teachers preparing their students for ELA questions in which they have to explain the main idea or write specific types of short essays (e.g. using the 'Hamburger method' of starting with an introduction, writing three supporting paragraphs and a conclusion where students are taught to include specific sentence starters and transition phrases). In Math all teachers express examples of how they use an instructional approach of student-led conversations to teach students to explain their thinking in math as they have to do on the state test. In some cases, teachers do a mini-lesson on fractions, then have students turn and talk about fractions; next they have to write down what they learned about fractions in a Math journal or on a weblog. Teachers also often use a distractor analysis, showing the percentage of students choosing specific (wrong) distractors, to explain to students why a specific choice of answer was incorrect.

A grade 2 teacher for example explains how she prepares students for the state test in grade 3:

> Let's say we have a reading, we have to read a passage, and we have to answer some questions, when we're doing test prep, sometimes what I do is I make, how do you call it, a transparency of a passage, and I show the whole class, everybody has a copy of the passage, and I model for them how I would answer questions in such a situation, so I read the passage and then I start looking at the question, and then I teach them, okay I answer my question, I make sure that what I said is in the passage, and when they sign their answers I make them underline their answers, and things that would help them when they have the real test.

Additionally a grade 4 teacher says:

> But essentially it is going to be, there are different kinds of questions and you just have to make sure you look at the questions, how to answer it and how to back it up. So we are going to start with a sort of say it out loud essay where we are going to talk it through and then they going to write it with a partner and the next day they are going to write another with a partner and by the end of the week they are going to write their own essay that has the same sort of structure. So this is exposing them to as many different essay questions that there might be because last year we didn't know, but this year we know that there is going to be a fiction and a nonfiction piece.

We also found schools extending their instructional block of ELA teaching to meet the changes in the ELA test where students are required to read for longer periods of time and answer more items. A teacher in grade 4 for example explains: 'We have been training all year, with read-aloud and reading the questions we have been preparing all year with the different genres. But right now we… for reading I mean I have a test prep, so for reading, one of the days, big focus is on just stamina, and one of the ways it is, just for reading 60 minutes straight. I think stamina is huge because I can't read 60 minutes straight without being distracted so I think that stamina is a big thing.' (school B).

Choice of assignments and formative assessments

Teachers in schools A, D, G, I and H explained how they align their own classroom assessments or formative tests (including rubrics for grading tests, content of tests, and format of tests) to the state test to familiarize students with the format of the state test and to learn how students would perform on the state test. Teachers created their own classroom assessments (quizzes, chapter tests, exit slips, do nows, etc.) that used rubrics, formats, or grading scales that were similar to the state test; they reviewed state test results to select the content that would go on their homemade classroom assessments; or they purposefully chose tests that were similar to the state test. Examples include the use of old state test items in classroom assessments, creating test questions that use the same grading rubric or format as the state test, and using the same proficiency levels as used on the state test in grading classroom tests.

Test preparation

All schools, except school A, schedule a period of approximately six weeks of test preparation before testing. During these weeks the schools focus on re-teaching the skills that will be on the test; in areas where students are failing, they have students practice old state tests or similar tests (using test prep material such as Kaplan books), or they teach students generic test taking skills such as filling in bubble sheets and eliminating and selecting the correct answer.

The above instructional activities are often enforced in the quality review when the reviewer addresses teachers' use of data to inform instruction or wen the quality reviewers explain to principals, assistant principals, math/ELA coaches and teachers how they should analyze the test and use item-skills analyses to differentiate and improve the rigor of instruction. They sit down with teachers and principals to analyze the ' tracking sheets' that are used by teachers and principals to monitor achievement levels and progress of students by class, grade, subgroups of students and for each subject, and they question the school about how instruction is adapted (e.g. through push-in and pull-out support of students or differentiating regular instruction) to ensure that students improve their proficiency level and the school meets AYP. The quality reviewers particularly look for the use of state test results and results on predictive assessments; sometimes even stating that they are not interested in the results on other internal assessments (school H). Some quality reviewers provide suggestions on the topics teachers should include in instruction and how to explain these topics to make sure students score higher on the state test; e.g. the

quality reviewer in school E and I emphasizes that teachers should focus on 'inferencing' to enable students to pass the ELA state test, to teach students to explain their answers to mathematical problems to pass the math state test or they motivate teachers to select 3-5 students in each class to target for additional monitoring.
Additionally, quality reviewers address a wide range of other instructional issues, such as teachers' use of rubrics to grade students' work and aligning these rubrics to the state test (School E, H), using measurable individual student goals (including targeted test scores) in instruction (school E), including math drills in instruction, using more challenging tasks in instruction (instead of rudimentary worksheets) (school C, F), instating information centers in classrooms (school C), differentiating instruction to modality and learning style of students (school C), using protocols to discuss students' work in teacher meetings (school C), use of computers in the classroom (school C), giving students more freedom in writing instead of drilling them to use specific structures of text (school A), or discussing the accuracy of a teacher's explanation of mathematical content (school F), etc..

Broad/narrow responses of multiple measures
Overall, adding quality reviews to the test-based accountability system in New York City seems to both broaden and narrow responses of schools. The quality review narrows responses of schools when schools pick up on feedback of the quality reviewer on how to narrow their goals, curriculum and/or instruction to improve students' performance on the state test. Examples of such narrow responses were found in schools D, A, H, I, E, F. The principal in school H for example explains how they will follow the reviewer's advice to only focus on improvement of student achievement on the state test in ELA and math (instead of also working on improvement of achievement in other subject areas), school E will specially target students on the cut point to a higher proficiency level and teachers explain to follow up on reviewer's feedback on how to use test items and scoring rubrics in the state test in their instruction and focus their instruction on tested skills.
The quality review also motivates broad responses of schools when they prepare for the review or act on the feedback of reviewers. Examples of a broadening of responses were found in schools C, D, E, F and H; the principal in school C for example expresses the intention to implement professional development of teachers on differentiation of instruction; school D plans on communicating curriculum standards to parents; school E intends to improve their progress reports on student achievement to parents; the principal and teachers in school F express the intention to implement a more structured social studies curriculum; and school H is working on action plans to improve teacher observations and progress reports to parents.
Schools also broaden their responses as a result of preparing for the quality review. Examples were found in schools A, C, D, E, F, H and I who start to improve the rigor and differentiation of their instruction; who start to implement individual student goals and have students reflect on their work; who implement common assessment rubrics; who improve writing samples of students; or who incorporate common core standards in the curriculum.

The feedback of the quality reviewer (his/her use of) the quality review rubric and the role of the quality reviewer seem to explain the extent to which this narrowing/broadening actually takes place.

Indicators in the quality review rubric
The quality review rubric includes five quality statements on 1) instructional and organizational coherence, 2) gathering and analyzing data, 3) planning and setting goals, 4) aligning capacity building and 5) monitoring student progress and revising plans and practices. Quality statements 2, 3 and 5 include indicators on how principals and teacher use (test) data, such as 'using collaborative and data informed processes to set measurable and differentiated learning goals for

students'. The line of questioning during the quality review also highlighted a strong focus on data use within the other two quality statements. Quality reviewers ask for example how teachers use data to inform and differentiate their instruction; they ask principals how they use data to track progress of school-wide goals; or they ask for data-driven goals for school improvement. Often, reviewers have analyzed the school's performance on the state test in detail and ask specific questions on how schools intend to improve performance of specific subgroups in specific areas. These include questions about low performance of Hispanic students (school A), performance of students on constructed response items in Math (school B) or performance of students that will get the school additional credit on the progress report (school H). Schools that target specific numbers or subgroups of students (to make progress or AYP) are complimented by the reviewer (e.g. school E), and observations of classrooms are sometimes targeted towards those that have low scores on the state test.
When quality reviewers gather evidence in schools to evaluate the school's functioning on the rubric, they also emphasize evidence that is related to the state test and to assessment data in general. One of the teachers for example explains that the quality reviewer assessed a lack of social studies curriculum in the school only because the school had no assessments and assessment data to show during the review.
The close connection between the quality review and the test is also felt by schools who have low grades on the progress report and feel that they will not get a well-developed score on the quality review. Some quality reviewers confirm this thought by expressing to a school with a C or D on their progress report that they have to make sure all the evidence is in place to give a proficient as the Department of Education will check the evidence very diligently in these cases. Another school (school B) refutes the critical remarks of the quality reviewer by pointing to their high test scores to legitimize their current (educational) practices.
The strong focus in the quality review rubric and the line of questioning and protocols for observations on data use in schools has the potential to narrow responses of schools to focus on improvement of student test scores.

The rubric in some cases also broadened responses of schools; particularly the first quality statement on instructional and organizational coherence and the fourth quality statement on aligning capacity-building motivated some schools to align their curriculum to the common core standards, to improve the rigor of instruction and motivate teachers to work in teams to evaluate and improve students' progress. Schools preparing for the review concluded that they didn't meet these indicators and started to improve in these areas.

Feedback
The feedback of the quality reviewer during the visit and in the written report is also an important explanation for narrow or broad responses of schools. Narrow responses occur when schools respond to feedback on having to improve their use of data to differentiate and target instruction to specific subgroups of students and /or tested skills. Quality reviewers in all schools provide some feedback on how to align curriculum and instruction to the state test and how to improve student performance on the state test (e.g. on improving data systems, aligning teaching to the test by increasing student writing in Math to increase scores on test items, implementing test prep for specific groups of students).

Providing feedback on the instructional quality of lessons or specifically observing and assessing instructional quality of lessons, other than how teachers use data to differentiate instruction, the rigor of lessons and engagement of students in the lessons, was not common practice during the observed reviews. Only in some cases did quality reviewers provide feedback on more 'in-depth' aspects of teaching to the principal (and not the teacher), for example when they observe teachers

providing an obviously wrong explanation of a math concept (e.g. school B and F) and when they have sufficient expertise to identify such instructional errors.
The quality review broadens responses in schools when the reviewer provides feedback on aspects of teaching and learning that are not specifically related to the state test. The feedback of quality reviewers of school B, C, D, E, F, G, H and I have the potential to broaden responses as they include suggestions on implementing a curriculum in social studies or improving the rigor of instruction.

Role of the quality reviewer
The feedback provided during the quality review is particularly important in broadening or narrowing responses of schools due to the double role of seven of the reviewers as superintendents for those same schools. This double role on the one hand increases the stakes for schools to act on the feedback (as the superintendent also evaluates performance of principals, approves the goals in the comprehensive education plan, decides on tenure of teachers and on promotion of students in testing grades); these superintendents also often express that the school's performance on the state test is important for their own evaluation and image as a superintendents. As a result, they target their observations of teachers to those with low test scores and they primarily question schools and teachers about teaching practices and action plans to improve student achievement results in ELA and math.

Superintendents who are reviewers also put additional pressure on schools by specifically referring to their authority during the review. In some cases, for instance, superintendents/reviewers said they would need to see a number of formal observations of teachers in specific areas (e.g. differentiation, rigor etc.) before they would give teachers tenure, or announced they would come back later in the year for a (superintendent) walkthrough of the school.
The double role of quality reviewers also motivates schools to act on the feedback as schools feel that their superintendent has a stake in high performance of the school on the state test and they trust the feedback is provided to help them achieve this goal. Two of the quality reviewers spok to these goals as they explicitly stated that they want the school to do well as their performance as superintendent is based in part on how their schools are performing on the State test; one of the two reviewers also explicitly stated that she wants to have the highest performing district in the city.

7. Conclusion and implications

Many studies point to potential unintended consequences of single measure accountability systems such as when schools narrow their teaching and instruction to fixate on tested subjects and topics and ignore important tested content. Several states and districts have therefore complemented the federal test-based accountability system with district or citywide inspections or quality reviews of schools to implement more effective multiple measure systems. These multiple measures are generally implemented to serve two (somewhat contrasting) purposes; in some cases they are expected to motivate schools to focus on a broader set of goals and to better connect to concerns such as student engagement in learning and instructional quality. On the other hand, multiple measures could also be implemented to improve the reliability of measuring a school's performance on a single indicator and to more effectively focus school improvement efforts, curricula, assessment and instruction to this single indicator.

Currently however it is not clear whether these multiple measure accountability models broaden school improvement or effectively focus schools' responses to improving a single indicator of student achievement. This study focused on the New York City accountability system which includes multiple measures (tests, quality reviews, school records, survey) to evaluate cognitive and non cognitive outcomes and educational practices of/in schools.

We expected the multiple measures of the New York City accountability system to broaden schools' responses to change their goals, curriculum and instruction in ways that go beyond what a single measure (e.g. the test) demands. Our results however reject this assumption as principals and teachers particularly refer to the state test when explaining their goals, curriculum and instruction. Schools in our study primarily set goals on improving student achievement (of specific subgroups of students) in ELA and math on the state test, and they align their curriculum and teaching to the test. Schools reallocate instructional resources to Math and ELA (discounting social studies and Science or other subjects such as Arts); teachers align their teaching in Math and ELA to topics that are tested, to when these topics are tested and to how extensively these topics are tested. They also use information on test items (e.g. item formats and scoring rubrics) to inform their instruction and coach students to do well on the test. They for example use test items in instruction, teach students how to explain their answer on Math items to get full credit on constructed response items, or include a specific number of paragraphs with introduction and transition sentences to get full credit on essay questions in ELA.

The quality review often reinforces schools' responses to the test, such as when principals for example follow the reviewer's advice to only focus on improvement of student achievement on the state test in ELA and math (instead of also working on improvement of achievement in other subject areas), or they specifically target students on the cut point to a higher proficiency level.

The quality reviews can however also broaden responses of schools. Examples are schools developing goals and activities to improve the curriculum in untested subjects, or to target and improve performance of all students (instead of only the students that will lead to quick gains in making AYP) when preparing for the review or acting on the feedback of the reviewer.

The feedback of the quality reviewer, (his/her use of) the quality review rubric and the role of the quality reviewer seem to explain the extent to which this narrowing/broadening actually takes place. The large focus on how schools use data in the five quality statements in the quality review rubric seems to narrow responses of schools to focus on improvement of student test scores. Quality reviewers give schools feedback on how to improve their use of data to differentiate and target instruction to specific subgroups of students and /or tested skills or they give suggestions on how to improve student performance on the state test (using test items or scoring rubrics).

The rubric and reviewer's feedback in some cases also broadened responses of schools, particularly the indicators on aligning the curriculum to the common core standards, the rigor of instruction and teachers working in teams to evaluate and improve students' progress lead to broad improvement actions. Schools preparing for the review concluded that they didn't meet these indicators and started to improve in these areas. In some cases, quality reviewers also provided feedback that broadened responses of schools; e.g. feedback on implementing a social studies curriculum (in addition to the curriculum in ELA and Math) or targeting improvement of all students (instead of only students on the cut point of a higher proficiency level).
The double role of the quality reviewer (often also being the superintendent of the school) increases the stakes schools have in acting on the feedback and additionally target their responses to improvement of test scores. Superintendents express that the school's performance on the state test is important for their own evaluation and image as a superintendent; superintendents also evaluate performance of principals, approve the school's goals in the comprehensive education plan, decide on tenure of teachers and on promotion of students in testing grades and therefore have additional means to make sure schools target their responses to improving their test scores and/or use the feedback provided during the quality review.

Overall, there seems to be a strong alignment of the quality review and the test in the focus on cognitive outcomes in both measures and in alignment of the grades (the quality review score and the progress report grade) awarded to the school. As a result, the quality review is transformed into another quantitative measure, focused on outcomes of schools, instead of functioning as a separate additional measure of educational practices in schools.

While the emphasis on aligning goals and work in the schools to the tests was unmistakable in these, several limitations of the study are worth noting. First, the study relies on nine case studies. The benefit of using case studies is to capture rich data on how schools respond to the different accountability measures and the specific circumstances that mediate such responses. As the topic of our study is relatively new, such an approach is opportune as there is no clear overview of the specific responses schools may choose that could inform a more quantitative approach to our data collection. At the same time, to develop these case studies, our data collection relies to a large extent on interviews in which teachers and principals retrospectively reflect on their decisions and actions. Such retrospective self-reports may bias our results as teachers may respond in a socially desirable way or they may not accurately remember why decisions were made about (changes in) the school's goals, curriculum and their instruction. Also the fact that observations were made during the quality review may introduce bias as the quality review is staged to assess schools; principals and teachers have a stake in showing positive responses to the accountability measures during the review. We tried to address these concerns by triangulating our interview data with data from other sources and only describing responses that were indicated in at least two out of the five data sources.
Another limitation is the fact that the data in these case study schools were only collected at two points in time (during the quality review and two to three weeks after the quality review). This time frame implies that responses occurring for example one month after the quality review or occur throughout the year are either not captured or only captured when teachers and principals describe them retrospectively. Also, the limited number of cases limits our results to an exploratory outline of potential responses to multiple measures that are by no means generalizable to all schools in New York City or to schools in other accountability systems.

Despite these limitations to generalizability, the cases illustrate the possibility that multiple measures, at least under some circumstances, may reinforce the narrow responses of single measure accountability systems that many think multiple measures should help to overcome. While some policymakers and administrators may see such reinforcement as a productive means

of focusing schools on outcomes that matters, in the process, measures like the quality review which could provide detailed information about processes and practices that might improve instruction become another measure of a school's outcomes. Thus, while contextualized descriptive information can be found in some of the school inspection reports in Europe (e.g. the Netherlands), such information is mostly lacking in the quality review reports in New York City where the feedback is limited to a score on a set of indicators on a 4-point scale and a set of relatively standardized sentences explaining these scores.

Furthermore, in this case, the use of multiple measures may exacerbate the use of educational practices that are designed solely to improve test scores, such as targeting students who are just below the levels that are needed in order for a school to make AYP. These practices will become increasingly relevant and valued, while other types of practices and goals (such as the ones on improving non-tested subjects) are ignored.

The quality review rubric, and the assumption that high performing schools on the test also perform high on the quality review, encourages schools to become more like what the rubric measures. When schools conform to the indicators in the quality review rubric, the rubric becomes a common, standardized and universal definition of which educational practices (e.g. the ones that lead to high scores on the state test) constitute good teaching. The institutionalization of such a definition encourages schools' behavior to conform to this definition and will increasingly narrow schools' responses to those that lead to increases of students' scores on the state test. As a result, what set out to be a multiple measure accountability system may, in the process of implementation and enactment, turn out to be a single measure system.

References

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, 85(1),* 7-38.

Booher-Jennings, J. (2005). Below the Bubble: 'Educational Triage' and the Texas Accountability System. *American Educational Research Journal,* 42(2), 231-268.

Chapman, C. (2001). Changing classrooms through inspection. *School leadership and management*, 21(1), 59-73.

Cullen, J.B. and Reback, R. (2006). Tinkering toward accolades: school gaming under a performance accountability system. Cambridge: National Bureau of Economic Research. Working paper 12286. http://www.nber.org/papers/w12286.

De Wolf, I.F., Janssens, F.J.G. (2005). *Effects and side effects of inspections and accountability in education; an overview of empirical studies.* http://www1.fee.uva.nl/scholar/wp/wp53-05.pdf

Ebrahim, A. (2005). Accountability Myopia: Losing Sight of Organizational Learning. *Nonprofit and Voluntary Sector Quarterly, 34(1),* 56-87.

Ehren, M.C.M. (2006). *Toezicht en schoolverbetering.* Delft: Uitgeverij Eburon

Figlio, D.N. and Getzler, L.S. (2002). Accountability, ability and disability: gaming the system. *NBER working paper 9307.* http://www.nber.org/papers/w9307.

Gong, B., & Hill, R. (2001, March). *Some considerations* of *multiple measures in assessment and school accountability.* Presentation 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.

Gribben, M.A., Campbell, H.L. and Mathew, J. (2008). Are Advanced Students Advancing? Examining Achievement Trends Beyond Proficiency. *Paper presented at AERA 2008.*

Haladyna, T.M., Nolen, S.B., Haas, N.S. (1991). Raising Standardized Achievement Test scores and the Origins of Test Score Pollution. *Educational Researcher, 20(5),* 2-7.

Hatch, T.C. and Honig, M.I. (2004). Crafting Coherence: How Schools Strategically Manage Multiple, External Demands. *Educational Researcher, 33(8),* pp. 16–30

Jacob, B.A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics,* 89(5-6), 761-796.

Jacob, B.A. and Levitt, S.D. (2003). Rotten Apples: an investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics (august)*, 843-877.

Jos, P.H. and Tompkins, M.E. (2004). The accountability paradox in an age of reinvention; the perennial problem of preserving character and judgment. *Administration & Society, 36(3),* 255-281.

Klerks, M. (in prep.). The effect of school inspections: a systematic review. *School Improvement*

Koretz, D.M. (2003). Using Multiple Measures to Address Perverse Incentives and Score Inflation. *Educational Measurement, 22(2),* 18-26.

Koretz, D.M., McCaffrey, D.F. and Hamilton, L.S. (2001). Towards a Framework for Validating Gains under High-Stakes Conditions. CRESST/Harvard Graduate School of Education: CSE Technical Report 551

Ladd, H.F. (2007). Holding Schools Accountable Revisited. 2007 Spencer Foundation Lecture in Education Policy and Management.
Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13(33),* 1-17.

Leithwood, K. and Earl, L. (2000). *Educational Accountability Effects: An International Perspective, 75(4),* 1-18.

Mehrens, W.A., and Kaminski, J. (1989). Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent? *Educational Measurement: Issues and practice, 8(1), 14-22.*

Mintrop, H. and Sunderman, G.L. (2009). Predictable Failure of Federal Sanctions-Driven Accountability for School Improvement-And Why We may retain it Anyway. *Educational Researcher, 38(5),* 353-364.

Nelson Espeland, W. and Stevens, M.L. (1998). Commensuration as a social process. *Annual Review Sociology, 24*, 313-343

Nelson Espeland, W. and Sauder, M. (2007). Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology. 113(1),* 1–40

Pallas, A.M. and Jennings, J.L. (2009). 'Progress' Reports (p.99-105). In: Ravitch, D., Meier, D., Avitia, D., Bloomfield, D.C., Brennan, J.F., Dukes, H.N., Haimson, L., Horowitz, E.m, Jennings, J.L., Koss, S., McAdoo, M., Ofer, U., Pallas, A.M., Sanders, S., Stern, S. Sulivan, P.J., Wolf, A. (2009). *NYC Schools Under Bloomberg and Klein: What Parents, Teachers and Policymakers Need to Know.* New York: Cass Size Matters.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. Educational Measurement: Issues and Practice, 10(4), 12-15.
Rosenthal, L. (2004). Do school inspections improve school quality? Ofsted inspections and school examination results in the UK. *Economics of Education Review,* 23(2), 143-152.

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 Hamilton, L.S., Stecher, B.M., Klein, S.P. (Eds.). *Making sense of Test-based Accountability in Education*. Santa Monica: Rand cooperation. http://www.rand.org/pubs/monograph_reports/MR1554/

Tetlock, P.E., Skitka, L. and Boettger, R. (1989). Social and Cognitive Strategies for Coping With Accountability: Conformity, Complexity and Bolstering. *Journal of Personality and Social Psychology, 57(4),* 632-640.

Visscher & R. Coe (Eds.), *School improvement through performance feedback* (41-75). Lisse: Swets & Zeitlinger.

Volante, L. (2004). Teaching to the Test: What Every Educator and Policy-maker Should Know. *Canadian Journal of Educational Administration and Policy,* (35), 1-6

Appendix.

Table 1. Background information on schools

	School A	**School B**	**School C**	**School D**
School profile: - Number of students - Subgroups	1371 students (preK-5), 10% Black, 11% Hispanic, 68% White, 6% Asian; 2% ELL, 14% special ed., 95.8% attendance	434 students (preK-8), 30% Black, 66% Hispanic, 3% White, 1% Asian, 50% ELL, 13% special ed. 91.8% attendance	437 students (preK-5), 70% Black, 28% Hispanic, 1% White, 1% Asian, 10% ELL, 13% special ed., 89.4% attendance	175 students (preK-5), 19% Black, 63% Hispanic, 10% White, 6% Asian, 5% ELL, 30% special ed., 92.1% attendance
Performance on measures	PR: A AYP for all subgroups and subjects Well-developed score on QR	PR: A AYP for all subgroups and subjects Proficient score on QR	PR: B AYP for all subgroups and subjects Proficient score on QR	PR: B AYP for all subgroups and subjects Persistently lowest achieving due to start of reporting Proficient score on QR

	School E	**School F**	**School G**	**School H**	**School I**
School profile: - Number of students - Subgroups	628 students (preK-5), 24% black, 72% Hispanic, 1% White, 1% Native American and 2% not reported, 15% ELL, 19% Special Ed., 91.5% attendance.	347 students (preK-5), 76% Black, 19% Hispanic, 02% White, 02% Asian, 1% American Indian students, 4% ELL, 19% special ed., 90.9% attendance.	1360 students (K-5), 23% Black, 72% Hispanic, < 1% White, 3% Asian, 33% ELL, < 1% special ed., 90.9% attendance.	544 students (preK-6), 73% Black, 18% Hispanic, 2% White, 4% Asian students, 9% ELL and 4% special ed., 91.6% attendance	470 students (K-5), 87% Black, 6% Hispanic, 4% White, 3% Asian students, 2% ELL, 10% special ed. 92.0% attendance
Performance on measures	PR: B Failed to make AYP in ELA for SWD, (status: in restructuring, advanced, focused) Proficient score on QR	PR: C AYP for all subgroups and subjects Proficient score on QR	PR: C School is 'Restructuring (advanced) - Comprehensive (ELA) for all students and all subgroups (student with disabilities are in safe harbor)	PR: D AYP for all subgroups and subjects School was on the state Persistently Lowest Achieving list and in danger of being closed	PR: D AYP for all subgroups and subjects Proficient score on QR

	School E	School F	School G	School H	School I
			and has a JIT in the same week as the QR School was on the SINI list last year Developing score on QR	Proficient score on QR	

[i] Everyday Math is a commercially produced Pre-K through 6th grade mathematics curriculum that has been adopted for use in many school districts.