Managing for Results at the New York City Department of Education

Citation

Childress, Stacey, Monica Higgins, Ann Ishimaru, and Sola Takahashi. 2011. Managing for Results at the New York City Department of Education. In Education reform in New York City: Ambitious change in the nation's most complex school system, ed. Jennifer A. O'Day, Catherine S. Bitter, and Louis M. Gomez, pg 87-108. Cambridge, MA: Harvard Education Publishing Group.

Permanent link

Terms of Use

Share Your Story

Managing for Results at the New York City Department of Education[1]

Stacey Childress, Harvard Business School
Monica Higgins, Harvard Graduate School of Education
Ann Ishimaru, Harvard Graduate School of Education
Sola Takahashi, Harvard Graduate School of Education

Paper Presented at the Annual Meeting of the American Educational Research Association
New Orleans, April 11, 2011

This paper was originally prepared for the November 2010 Conference of the New York City Education
Reform Retrospective Project and will be published in a slightly edited form in
Education Reform in New York City:
Ambitious Change in the Nation's Most Complex School System,
scheduled for release in spring 2011

Managing for results can be a treacherous tightrope walk for school districts. The popular adage calling for a combination of "pressure and support" remains easier said than done.[2] When schools are failing children, the natural impulse is to ratchet up accountability, adopt a "no excuses" policy of one kind or another, and increase the pressure on those closest to the students —principals and teachers. However, like most performance appraisal systems, if implemented in isolation this approach runs the risk of crowding out capacity-building just when it is needed most. Indeed, the schools that are penalized for low performance are often the ones staffed with the educators most in need of development. Though some schools might have a critical mass of staff who simply do not believe improvements in their practice would have an effect on student learning and therefore resist accountability, most reformers operate on the assumption that the vast majority of educators are motivated to serve their students well. As one long-time urban superintendent says, "If low-performing schools knew how to fix themselves, they would do it."[3] This implies that simply adding more pressure on schools without a concurrent effort to increase capacity runs the risk of undermining the ultimate goal of improving performance; with low self-efficacy or a lack of confidence in their ability to successfully adopt new approaches, principals' and teachers' prospects of improving their students' learning may be hindered. At worst, these professionals might engage in undesirable strategic behaviors that undermine learning.[4]

How then can school districts walk the high wire in service of improving results for all children? The New York City Department of Education (DOE) has an approach that has sparked interest from policy-makers, practitioners, and researchers worldwide. This paper attempts to describe this approach. We begin by describing the history behind the DOE's evolving theory of action for how to improve student performance at scale. We include a summary of their efforts to, as Chancellor Joel Klein often said, "change the facts on the ground" to support implementation of the theory between 2002 and 2009. To do this, we cover some of the ground in Paul Hill's chapter in this volume, but focus more specifically on the strategic and structural choices that are relevant to the performance management system. As we will show, the perspective that the DOE took on performance management was unique at the time: rather than granting autonomy as a reward when schools achieve high performance, the DOE granted schools autonomy outright in return for high levels of accountability.

Following this historical overview, we offer a brief summary of research on "organizational learning," an idea that is reflected in the ways in which the DOE attempted to build capacity in schools. Next, building on this historical and theoretical backdrop, we describe the details of the DOE's performance management system and the ways in which schools have been held accountable while at the same time being provided with a variety of supports to do their work. Here, our chapter complements others in this volume, including Joan Talbert's examination of inquiry teams, an organizational learning mechanism that has become central to the DOE's reform efforts. Finally, we offer insight from some exploratory analyses we performed based on a large-scale survey of New York City teachers to investigate the DOE's theory of performance management in action. Here, we investigate teacher perceptions of school culture along the dimensions of accountability and organizational learning and explore whether and how these aspects of a school's culture relate to school performance. Suggestions for future research and implications for practice conclude our paper.

NEW YORK'S THEORY OF ACTION AND RELATED ORGANIZATIONAL CHANGES[5]

Between 2002 and 2009, Chancellor Joel Klein and his team implemented an evolving theory of action at the DOE based largely on the premise that if given autonomy for decisions about resources and instruction, school principals would improve the performance of their students more effectively than if those decisions were made for them at the central office. In exchange for this autonomy, principals and

their schools would be held more accountable for results than in previous administrations. Klein's theory development was heavily influenced by the work of scholars and thought leaders such as William Ouchi, Paul Hill, and Michael Barber, as well as by his observations of a number of high-performing charter schools in the city.[6]

In order to put their theory into practice, Klein and his team implemented a two-phased strategy. In the first phase (roughly 2003–2006), the team sought to consolidate control by moving the oversight of New York's approximately 1,300 schools from 32 separately governed and managed community districts and a high school division to 10 K-12 regions, a special education district, and a district for alternative schools and programs (for a total of 12 entities), all reporting to the chancellor and governed by Mayor Michael Bloomberg. These new structures were designed to support schools' operational and instructional needs, enforce standards, and implement reforms. Dubbed *Children First*, this first phase also instituted a common math and literacy curriculum for grades K–8, ended social promotion, created 150 small schools to replace large low-performing high schools, and added a parent coordinator and math and literacy coaches to every elementary school. Reflecting on this phase in 2007, Klein said the purpose was to gain control of a "chaotic and dysfunctional organizational structure" and to "lock the system down, establish some control, and bring coherence to the system." Required curriculum, standardized promotion policies, and a more centralized reporting structure seem incongruous with a theory of action premised on greater autonomy. Klein and his team have argued that in order to eventually put more control in the hands of school leaders they had to first disrupt the entrenched bases of power in the 32 community districts, and transferring those districts' authority to the new regions was one way to accomplish this.

After the first two years of Children First, Klein and the senior team began laying the groundwork for phase two, which would eventually turn the system on its head by moving power from the 12 entities to the leaders of nearly 1,300 schools.[7] Klein believed that empowering highly competent principals and giving them the resources needed to make decisions about solving their schools' performance problems would ultimately be in the best interest of students. In order for empowerment to be effective, Klein also believed that principals needed to be held accountable for student performance. Klein described his theory of action for phase two, saying:

> If we empower principals and hold them accountable for school results, we'll do two things—shift the locus of power from central office to the schools and shift the organizational culture to a focus on results. However, I know that autonomy in and of itself is not going to guarantee success. But it will lead to innovation. And I suspect that if we're tight on accountability and instill an intense focus on student outcomes, we can also build into the equation some variability in terms of problem solving at the school level and learn from it.

The first wave of phase two was a pilot called the "autonomy zone," which school leaders could opt into by signing performance contracts in exchange for freedom from many of the mandates of the Children First reforms, including their reporting relationship to a regional superintendent and the district-wide math and reading curriculum. Other large urban districts, including Chicago, Boston, and Philadelphia, were experimenting with such zones, but in those cities, schools earned autonomy by meeting performance thresholds. In New York, the senior team developed a fundamentally different assumption: autonomy was not something to be earned by performing well; rather, it was a prerequisite for high performance. Any school leader could choose to participate in the autonomy zone (the Zone), regardless of the school's current performance.

Twenty-nine schools volunteered to join the Zone, more than half of which had recently launched as new small schools. Each principal signed a five-year performance contract that specified targets for a variety of indicators, including state test scores, attendance figures, and graduation rates. As part of the agreement, principals assumed control over all budgetary and decision-making authority that previously resided with regional offices. Zone schools that met their performance targets were left alone, while those that missed more than a certain number of targets entered into what the DOE called the "ladder of consequences." The timeline for these consequences was more aggressive than those in No Child Left Behind or the New York state accountability system. The first year a school missed its targets, the school's leadership team was required to develop specific action plans for the following year. If a school missed its performance goals for a second year, the principal could be removed. If things did not improve by the third year, the school could be closed. The principals answered to the Zone chief executive officer, Eric Nadelstern, who was responsible for all Zone schools and worked directly with those who failed to meet their targets. At the beginning of the second year, 19 additional schools joined the Zone. At the end of the second year, 2 of the 48 Zone schools failed to meet their performance goals. In fall of 2006 the number of schools in the renamed "empowerment zone" grew to 332. Klein heralded the increasing interest in the autonomy/accountability exchange as a shift in the fundamental culture of the school system, saying,

> I think if you don't change the culture of public education, you're not going to change the outcomes materially. A culture that doesn't focus on performance is a culture that won't work. And now these principals are saying they want to be held accountable for their performance. It's quite a thing to have [332] principals, over the objection of their union, sign performance agreements saying, 'If I don't hit the ball, they'll be sending me to the minors.' It's a big, big deal given the objections from the union and the good relationships many of these principals have with their regions.

Building on what they learned in the pilot, the senior team unveiled a new three-pronged organizational structure that went into effect in fall 2007. The special education district remained intact, but the regions were disbanded and every school leader was given the choice to affiliate with one of three types of school support organizations (SSOs):

1. **Empowerment Support Organization (ESO)**: The empowerment zone converted into the ESO, and additional schools could become empowerment schools by signing performance contracts and forming self-managed networks.

2. **Learning Support Organizations (LSOs)**: Most similar to the old regional structure but not organized geographically, four learning support organizations offered schools service, support, and oversight through four centers run by the DOE.

3. **Partnership Support Organizations (PSOs)**: Schools that partnered with external nonprofit organizations for support and services were part of this structure and signed performance contracts. A number of schools were already partnered with external organizations, and the new structure would formally give those organizations additional support responsibilities, hold them accountable for school results, and open up the option to additional schools and nonprofits. PSOs had the same level of autonomy granted to the school operators in the ESO.

The autonomy/accountability exchange was at work regardless of which support organization school leaders chose. Though the ESO included the most freedoms, principals' choices about which support organization to affiliate with were framed as acts of autonomy, and regardless of which choice they made, every principal signed a five-year performance contract. Principals also had the option to transfer

to another support organization during an open enrollment period every two years. Klein's team believed this new level of autonomy would attract new people to school leadership positions who would be more likely to innovate.

As part of the autonomy equation, Klein and Bloomberg negotiated mutual consent hiring with the United Federation of Teachers (UFT), the largest local bargaining unit in the United States. Prior to this agreement, tenured teachers were able to choose which schools they worked in regardless of whether a school's principal agreed to the transfer. If no openings existed in the desired school, a senior teacher could "bump" a non-tenured teacher out of the school to create a slot. Under mutual consent, principals became responsible for hiring their teaching staffs and "bumping rights" were eliminated. Klein considered this change a pillar of the autonomy/accountability exchange—if principals were to be held accountable for the commitments they made in their performance contracts, they needed the flexibility to build school teams they felt best positioned them for success.

In order to make the autonomy/accountability exchange work at scale, a set of coherent systems and structures were necessary to capture the benefits of principal autonomy and at the same time exploit the advantages of being a large, integrated system. The next section describes some of the key accountability and organizational learning mechanisms that together make up the performance management system in New York City.

DESIGNING A PERFORMANCE MANAGEMENT SYSTEM

Jim Liebman had been a respected civil rights attorney, legal scholar, and teacher at Columbia Law School before joining the DOE in 2005 as the chief accountability officer. Klein recruited him with a mandate to lead the design and implementation of a performance management system. Liebman often described the performance management system his team created as an attempt to strike the right balance between instilling accountability for past results and encouraging continuous improvement toward future outcomes. A key goal of the system was to help schools develop into organizations in which professionals were constantly learning with one another about how to solve performance problems. Liebman described his thinking this way:

> Accountability isn't entirely or even mainly about incentives. It's about capacity building, which to me means adult learning based on self and team evaluation of what's working and what's not, and knowledge management, meaning spreading what works from one student or school to another. If we want the lever of accountability to be as powerful as possible, we have to provide ways for schools to build their capacity to be relatively self-sufficient in evaluating themselves every day and in solving their unique performance problems and, when necessary, in asking for the specific help they need. This will never work if the central bureaucracy behaves as if it has all the answers. Our role is to help professionals in schools ask better questions so that they can craft customized answers based on their own evaluation of their performance problems.

Organizational Learning and Accountability: A Theoretical Backdrop

The DOE's approach to capacity-building is consistent with relatively new approaches to school reform that include the idea of organizational learning. Rather than providing training to implement a specific curriculum or instructional technique, the DOE focused on increasing school teams' capacity to learn from data and adjust their behavior based on this learning. "Organizational learning" refers to a collective or a more macro-level learning than that which occurs at the individual level.[8] There are many excellent reviews of the organizational learning literature and its application to the context of education,

including a recent 2008 special issue on organizational learning in the *American Journal of Education*.[9] Rather than provide yet another review, we offer instead a very brief overview of the literature, highlighting specific aspects of this work that are present in the design of the DOE's performance management system.

In general, we note that despite the espoused merits of bringing organizational learning theory to the field of education, many education scholars concur that empirical evidence in this arena is lacking and that explicit approaches to building capacity for organizational learning in schools remain in their infancy. One reason that this has remained an elusive phenomenon may be that the concept of organizational learning has become burdened with multiple interpretations. The term "organizational learning" was originally introduced by March and Simon in 1958.[10] Their research, along with that of their doctoral students, focused primarily on the cognitive processes of information search, acquisition, integration, and assimilation. Much of this work was on private sector organizations, and centered on how knowledge and information could be resources to help organizations "learn" and thereby improve performance.[11]

In contrast, a second approach to organizational learning has emerged over the years that scholars such as Meredith Honig describe as "sociocultural learning theory."[12] Here, rather than emphasize the rational, cognitive processes that govern information management and complex problem-solving, the focus has been on how individuals co-construct meaning with those around them. From this vantage point, learning cannot be considered without a close examination of the social embeddedness of individuals in organizations. This research stream has focused on how individuals learn through interactions with others in practice and on the social processes of learning.[13;14]

Recently, education scholars have called for greater integration between these two perspectives on organizational learning.[15] Research in the adjacent field of organizational behavior has adopted a perspective that takes such an integrative stance. In 2008, Garvin, Edmondson, and Gino introduced several "building blocks" that they argue are critical to organizational learning: a supportive learning environment in which individuals feel "psychologically safe" to speak up and ask for help; concrete learning processes and practices that enable the collection, analysis, and transfer of valuable information; and leadership that reinforces learning by supporting those closest to the organization's core work.[16;17;18] Although much of these scholars' work has focused on healthcare and other industries, the conceptual underpinnings of their research and the integrative stance they take has caught the attention of reformers such as Liebman.

Today, many scholars and practitioners, like Liebman, view organizational learning as one side of a productive tension. The other side is accountability. Since the advent of standards-based reform in the 1990s, scholars have been writing about its implications for accountability. This research stream has definitional and interpretation challenges similar to the organizational learning literature cited above. Jennifer O'Day's work draws a distinction between bureaucratic and professional accountability that helps to distinguish between the often top-down pressure an organizational system puts on educators and the mutual responsibility educators have to one another to live up to the quality standards of their profession.[19] Richard Elmore categorizes the different dynamics as internal accountability, the shared expectations for teaching and learning and a means to meet these expectations; and external accountability, a system for holding people responsible for producing certain results.[20] Elmore also writes about the notion of reciprocal accountability, which he sums up in the following statement: "For every unit of performance I demand of you, I have an equal responsibility to provide you with a unit of capacity, if the performance I require of you requires knowledge and skill that you do not possess."[21]

Bringing multiple research streams together, organizational behavior scholar Amy Edmondson recently proposed that work environments characterized by high levels of organizational learning as well as high levels of accountability enable superior performance.[22] This proposition has yet to be empirically tested. Still, it echoes the present-day thesis among education leaders that it is important to balance "support and pressure" in order to reach high performance in schools and district-wide. When communicating the DOE's performance management tools to internal and external audiences in the early days of implementation, Liebman would often quote Elmore's reciprocal accountability concept, and the particular form of capacity he emphasized was the ability of school teams to learn together how to solve their performance challenges. This framing is consistent with Edmondson's proposition that high levels of both accountability for results *and* organizational learning will lead to higher performance than the presence of only one or the other.

IMPLEMENTING SYSTEMS FOR ACCOUNTABILITY AND ORGANIZATIONAL LEARNING

Consistent with the theoretical frame described in the previous section, Liebman and his team in the Office of Accountability developed learning tools and processes to help schools build their capacity to problem-solve and self-evaluate, even as they ratcheted up the accountability pressure for results. The organizational learning and accountability tools together formed a performance management system with a number of key components, including school progress reports, quality reviews, periodic assessments, inquiry teams, and a new technology system.

In this section, we focus specifically on the design, implementation, and evolution of progress reports as an accountability tool and quality reviews as an organizational learning tool, as well as some of the interactions between them. Other chapters in this volume discuss other aspects of the performance management system (see, for example, Joan Talbert's chapter on inquiry teams). We primarily cover the first three school years of implementation (2007, 2008, and 2009) and only refer to changes implemented in the 2010 school year when they are useful in illuminating challenges encountered in the preceding years.

Progress Reports

The Office of Accountability created a progress report to evaluate and communicate school performance internally and externally. The reports were designed to provide a historical account of a school's overall success in improving student academic outcomes. Progress reports for all schools included four main subsections, each with different weights that would add up to 100 points to determine a school's overall letter grade, as discussed in more detail below.

To account for differences in curriculum and targets by grade level, two versions of the report were created, one for elementary and middle schools and another for high schools. In 2006–2007, letter grades were assigned based on how schools performed relative to one another on the 100 point scale: The top 15% received As, the next 40% received Bs, the next 30% received Cs, the next 10% received Ds, and the bottom 5% received Fs. After a baseline was established the first year, each September the DOE published the cut scores that would correspond to letter grades on the progress reports at the end of the year.

The cut score for various grades increased a few points each year in order to continuously raise the bar for system-wide performance. The goal was to push out the performance horizon so that schools with good letter grades had incentives for continuous improvement.

The following sections describe the four basic subsections of the progress reports: school environment, school performance, school progress, and additional credit; and how they are calculated to produce a score on the 100-point scale and a letter grade.

School environment

The school environment section evaluated schools based on attendance figures and survey data from parents, teachers, and students in 6th through 12th grades. The 15% weight for this subsection was the sum of 5% for attendance and 10% for the survey results. The surveys covered four areas, each of which contributed equally to the 10% weight: safety and respect, academic expectations, engagement, and communication. Liebman's team designed the surveys but a vendor delivered them, collected the data, and provided anonymous results back to the DOE. Administering approximately 1.5 million surveys posed significant implementation challenges, but from the first year the response rates were large enough to allow for reliable data, and the response rates grew each year, as shown in Table 1.

Table 1: Environment survey response rates

	2007	2008	2009
Teachers	44%	61%	73%
Students (6–12)	65%	78%	80%
Parents	26%	40%	45%

Source: Author analysis of publicly available progress report results for 2007, 2008, and 2009.[23]

Student performance

The student performance section graded the school based on the percentage of students scoring at level 3 or 4 on that year's state mathematics and English-language arts exams. On the high school progress report, this section scored performance on Regents exams and graduation. In 2007, this subsection was worth 30% of the total score, and was reduced to 25% in subsequent years.

Student progress

To distinguish the new progress reports from previous accountability reports distributed by the city and state, the DOE emphasized individual student growth from one year to the next much more than the average absolute performance of a school's students at a point in time. This section started with a weight of 55% and increased to 60% after the first year of implementation. Beginning in 2008, for elementary and middle schools the progress section measured the change in individual students' performance with four indicators:

- Percentage of all students making at least one year of progress on state exams

- Percentage of students in the school's lowest third making at least one year of progress on state exams

- Average change in proficiency for students scoring at level 1 or 2 in prior year

- Average change in proficiency for students scoring at level 3 or 4 in prior year

Instead of rewarding schools for attracting students who already performed well, the DOE hoped the heavy weight of this subsection would reward what schools brought to their students—in other words, giving schools credit for "adding value" to their students' performance trajectory. In order to account for movement within a proficiency level, the DOE converted the state's range of raw scores to a proficiency rating for ELA and math from third through eighth grades. Table 2 includes illustrative conversions for eighth grade ELA and math scores.

Table 2: Conversion of New York state proficiency levels and raw scores to DOE proficiency ratings, 2009

Eighth Grade English Language Arts			Eighth Grade Math		
Level	Raw Score	Proficiency Rating	Level	Raw Score	Proficiency Rating
4	715-795	4.00 – 4.50	4	701-800	4.00 – 4.50
3*	650-714	3.00 – 3.99	3*	650-700	3.00 – 3.99
2	602-649	2.00 – 2.99	2	616-649	2.00 – 2.99
1	430-601	1.00 – 1.99	1	470-615	1.00 – 1.99

Note: *Level 3 is considered *proficient* for NCLB purposes
Source: NYCDOE Office of Accountability

For example, if a student started the school year with a proficiency rating of 2.25 and ended the year at 2.50, the school would receive credit for the gain even though the student did not advance from level 2 to level 3. On the flip side, a school would lose points if a student's proficiency rating went down. If a student's proficiency rating stayed the same from one year to the next, that was counted as making one year of progress. To account for student mobility within the city, a transfer student's progress was split between the sending and receiving school in the year of the transfer, based on the time of year the state exams were administered. A student's ELA progress was allocated 60% to the sending school and 40% to the receiving school, while math progress was allocated 40% to the sending school and 60% to the receiving school.

For high schools, the student progress section was calculated based on how students performed relative to the probability they would pass or fail the Regents exams given their eighth-grade scores on state standardized tests. This section also scored the average pass rate for Regents exams and credit accumulation. Liebman included credit accumulation based on historical data that showed a strong correlation between the number of credits a student had in specific semesters (especially in ninth and tenth grades) and the probability of graduating on time. Student progress was calculated on a semester basis in order to account for mobility—the student's progress was used to calculate the score of the school awarding credit in a particular semester.

Closing the achievement gap

The achievement gap section awarded additional credit to elementary and middle schools that raised proficiency levels by half a level or more for students in the lowest third of performance city-wide. The comparable section on the high school version gave additional credit to schools based on the number of students in the same group that attained at least a quarter of the credits needed to graduate in each of their first three years of high school.

Overall grade

In all but the additional credit subsection, schools received two scores: one showing how well the school performed in relation to all NYC public schools, and another demonstrating how the school performed relative to a group of 40 peer schools with comparable students. No school had the same 40 peer schools. To determine the overall grade, the city and peer comparisons for each subsection were weighted 25% and 75%, respectively, to derive a total subsection score (which was also converted into a subsection letter grade). The subsection scores were then summed for a total score on the 100-point scale. Any points from the additional credit subsection were then added and the resulting score was converted into a letter grade as shown in Table 2. **(**For a technical explanation of the calculation of each subsection and the overall grade, see the New York City Progress Report Educator Guide, available for elementary and middle schools as well as high schools on the DOE's website.[24])

Challenges: Fairness, Fit and Reliability

The progress reports quickly became the primary vehicle for assessing and communicating school performance. Even as they were embraced around the city as an accountability tool, their implementation generated a number of challenges related to their fairness, fit with the state's accountability system, and reliability.

Questions of fairness

Opening new schools was a major pillar of Klein's improvement strategy, and the DOE opened hundreds between 2006 and 2009. New elementary and middle schools did not receive a progress report until they had two years of data and high schools did not receive them until their first cohort graduated. This meant that the new schools strategy went largely "ungraded" in the first years of the progress reports. In addition, various stakeholder groups asserted that the progress reports did not account for special situations well enough. The DOE eventually responded to these concerns by developing separate reports for schools serving only K–3 students in which only third graders took state exams, for schools serving only special-education students taking alternative assessments, and for high schools serving students who were at risk of "aging out" of the school system without graduating.

Some school communities argued that they were disadvantaged in an evaluation system with a 60% weight on student growth because they already had high levels of student achievement. But the DOE leadership team was firm in its belief that schools should be rewarded for what they added to their students, and resisted calls to create a customized version of the report for the city's perennially high performing schools. However, a new rule would go into effect in 2010 so that no school in the top quartile of the performance subsection citywide could receive a progress report grade below C regardless of their score on the progress subsection.

Fit with state accountability system

The DOE's timeline for imposing consequences on underperforming schools was more aggressive than the state's timeline, but the state evaluated schools on absolute performance, not growth rates. As a result, schools with low absolute performance and rapid growth rates received As and Bs on their DOE progress reports but showed up on the state's watch list for underperforming schools. This mismatch contributed to public confusion about how to assess school quality.

Another tension was related to the state's annual decision about the cut scores used to define proficiency levels on the state exams. For instance, the state lowered the cut scores for each level in 2009, making it much easier for students to reach the proficiency bar. Because of the heavy weight on year-over-year growth in the progress reports, 84% of schools received an A or B on that year, as shown in Table 3. The dramatic uptick in letter grades called into question the usefulness of the progress report in determining the real state of performance in the city's schools.

Table 3: Percentage of schools with each progress report grade, 2007–2009

	A	B	C	D	F	No score
2007 (n=1261)*	23%	39%	26%	8%	4%	n/a
2008 (n=1493)	35%	37%	13%	4%	2%	9%
2009 (n=1527)	68%	16%	5%	2%	<1%	8%

*In 2007 the DOE did not include schools without scores in its progress report data file
Source: Author analysis of publicly available progress report results for 2007, 2008, and 2009.[25]

By fall of 2009 when the scores were public, Shael Suransky had succeeded Jim Liebman as head of the accountability office. He posited that if the leadership team had been able to predict in September that such a large jump might occur, they might have raised their own cut scores even higher so that the letter grades would have been a more accurate reflection of the relative performance of schools.[26] However, he also noted that the DOE would have had to raise the cut scores more than a standard deviation from the year before, which would most likely have created questions about the fairness of the decision.[27]

Concerns about reliability

Perhaps the most significant criticism came from measurement experts such as Harvard professor Dan Koretz, who pointed out that New York state exams were designed to provide a snapshot at a point in time, not as a way to track individual student growth from year to year.[28] Therefore, they were not appropriate for the task of measuring progress at the school level. Another criticism was that the unpredictability of changes in student scores from year to year suggested high levels of measurement error in the state exams, which accounted for 60% of the progress report grades in elementary and middle schools. High school reports were based on a more diverse set of measures and therefore were more stable.

After years of analysis and discussion of this issue, as well as the concern that those schools with high absolute scores were disadvantaged in the scoring system, the DOE announced that for the 2010 school year it would discard the four indicators it used to calculate the progress subsection in favor of a growth percentile for every student. While maintaining that the four indicators used from 2007 to 2009 had created an appropriate incentive for schools to maximize the growth of every student, the DOE acknowledged that they did not fully account for the relationship between the performance and progress metrics. Going forward, the progress section would use only two indicators: the median growth percentile for all students and the median growth percentile for the bottom third of students on the prior year's proficiency rating.

The new method would account for the fact that low performing students were more likely to show one year of progress on state exams than high performing students whose absolute scores were already at the top of the range. The approach was modeled on a methodology developed by the state of Colorado and took into account students' beginning proficiency levels when calculating their yearly progress for

state and federal accountability purposes. The underlying assumption was that growth percentiles would control for the probability that students would make one year of progress given their starting proficiency. The percentile would indicate what percentage of students had less growth than an individual student. For instance, a student who had a 3.0 proficiency rating on the fifth grade exams would be in the 35th percentile if he or she maintained a 3.0 rating on the sixth grade exams. A student who maintained a 4.2 rating from year to year, however, would be in the 50th percentile, because fewer students at that rating had historically made a year's worth of progress the following year. In the past, these two students would have contributed the same amount to a school's progress grade. In the new system, the student who began at a higher level would contribute more because of the higher percentile ranking.

In July 2010, as we were writing this chapter, the state of New York released exam scores for the prior school year, which had been calculated based on a new, much higher cut score for proficiency. As a result, scores around the state dropped dramatically. New York City's scores fell at a lower rate than the rest of the state, but nevertheless were a significant drop from 2009. Suransky and his team were recalibrating all scores back to 2006 based on the 2010 cut scores in order to create a better picture of what the growth rates had actually been over the years and were analyzing the implications the sudden change at the state level would have on progress report calculations.

Quality Reviews

The leadership team described the progress report as a *lagging* indicator because it provided a snapshot of how students performed relative to the year before and how schools scored relative to other schools in the city at a point in time. However, school teams needed 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. For this task, the Office of Accountability designed a quality review process to communicate and reinforce a set of behaviors and practices it believed would drive improvements in student achievement.

Beginning in the 2007 school year, every public school in New York City participated in an annual quality review (QR), modeled loosely on public school inspections in the U.K. and Hong Kong. Every school received a one- or two-day visit from an outside reviewer who observed staff meetings, examined documents, and asked teachers, parents, and the principal a series of questions about a number of the school's practices.

The Office of Accountability partnered with Cambridge Education, a United Kingdom-based school review firm, on the design and implementation of the QR protocols for three years. The rubric reviewers used to observe schools was focused primarily on the school's use of data to adjust teaching practices in order to improve student outcomes. Liebman and his team created five overarching quality statements, each of which had seven indicators that guided reviewers' observations:

- **Gather Data**: School leaders consistently gather and generate data and use it to understand what each student knows and can do, and to monitor students' progress over time.

- **Plan and Set Goals**: School leaders and faculty consistently use data to understand each student's next learning steps and to set suitably high goals for accelerating each student's learning.

- **Align Instructional Strategy to Goals**: The school aligns its academic work, strategic decisions, and resources—and effectively engages students—around its plans and goals for accelerating student learning.

- **Align Capacity Building to Goals**: The development of leadership, teachers and other staff capacity is aligned to the school's collaboratively established goals for accelerating the learning of each student.

- **Monitor and Revise**: The school has structures for monitoring and evaluating each student's progress throughout the year and for flexibly adapting plans and practices to meet its goals for accelerating learning.

Because the rubric focused heavily on the process by which school teams used data to improve instruction, it reinforced other organizational learning tools that are not discussed in this paper. These include the periodic assessments that every student took five times a year to map their progress toward mastery of the state standards, as well as the teacher-led inquiry teams that were established to spread a culture of data use throughout the schools. Liebman explained the rationale behind a rubric focused on the use of data:

> I looked at all of the rubrics that are out there for school reviews and instructional walkthroughs. They all have one section on the use of data to drive strategy and facilitate frequent adjustment, and many other sections on inputs such as specific teaching practices and curricular materials. As I thought about what we are trying to accomplish, I realized that what we needed to change most was the emphasis in our schools from 'teaching inputs' to 'learning outcomes.' Until we instill a culture of data-driven instructional differentiation in our schools, I'm not as interested in the parts of these rubrics that are about inputs. Secondly, we are not pushing a particular educational philosophy or professional development strategy. The last thing we want to do is to convey the sense that we have an idea in our heads at the central office about what the 'right' answer is for each school."

Liebman sometimes described the rubric as a way to determine the degree to which schools were on the way to becoming "autonomous problem-solving units."

In spring 2006, the DOE worked with Cambridge Education to pilot the process in 100 schools, which led to modest changes to the rubric and review process based on feedback from principals and reviewers. For instance, the pilot used three rating categories: underdeveloped, proficient, and well developed. Pilot schools overwhelmingly scored in the middle category, but principals and reviewers reported that they had seen a wide range of practices within the category. As a result, the Office of Accountability expanded it to five rating categories:

- Underdeveloped
- Underdeveloped with proficient features
- Proficient
- Well developed
- Outstanding

Before going through the quality review process, every principal went through approximately 40 hours of training on the process through a program called Children First Intensive (CFI) developed and run by

Irma Zardoya. CFI familiarized principals with the QR process, the five quality indicators and 35 sub-indicators, and the associated team behaviors reviewers would look for during the QRs.

Throughout the 2007 school year, Cambridge Education consultants performed all of the quality reviews, but a number of New York principals and administrators observed reviews so that they could conduct the process without consultants in the future. By June 2007, approximately 140 consultants had conducted school-quality reviews across more than 1,400 schools; at the end of 2008 they had conducted 2,900 reviews over the two years.

The process was costly; the three-year Cambridge contract was $19 million[29] and schools spent an enormous amount of time prepping for QRs every year. Beginning in 2009, the DOE no longer reviewed every school annually. Schools that received an A or B on the progress report combined with a QR rating of proficient or better were only scheduled for review every third year as long they maintained or improved their letter grade. Schools with other grade and QR combinations remained on the annual cycle. In addition to freeing up time at the school level, the reduction in the number of schools reviewed annually meant that the DOE could accomplish all QRs with internal reviewers rather than expensive outside consultants. At the same time, the DOE made changes to some of the sub-indicators, tightening some general language to more concrete statements of behaviors. Because only the lowest performing schools now received QRs and they were conducted by insiders who were on the lookout for more specific behaviors, the process felt much more compliance oriented than the DOE intended. Suransky and his team were considering ways to counter this perception.

Interactions between the QR and Progress Report

The progress reports and quality reviews were independent tools; for instance, the quality review rating appeared on the progress report but was not used in the calculation of the school's letter grade. However, Liebman and his team believed that viewed together they provided a picture of how the school had performed in the past and how it was likely to perform in the future. This could be useful in making decisions about positive and negative consequences for schools. Table 4 is a "consequences matrix" based on schools' progress report grades and QR ratings.

Table 4: Consequences matrix

		Quality Review Rating				
		Outstanding	Well-Developed	Proficient	Underdeveloped w/ Proficient Features	Underdeveloped
Progress Report Grade	A	Eligible for Rewards		No rewards or consequences		
	B					
	C	No Rewards or consequences unless school receives C for three consecutive years, then treated as a D school				
	D	Subject to academic planning and target setting by Office of Accountability.				
	F	Multiple years of D or F grade and no improvement in QR rating could lead to leadership change or school closure		Possible leadership change or school closure		

Source: Internal NYCDOE documents and primary interviews

In practice, the matrix was not used as an absolute determinant of a school's future, but rather helped identify schools for further attention from the chancellor and the accountability office. For example, a principal in a school that received both an F and an *underdeveloped with proficient features* rating was not automatically fired, but instead was flagged so that the senior team could examine other relevant information about the principal and school in order to make a decision about the school's future.

Plotting schools on the consequences matrix as in Table 5 provides an additional layer of information about how schools with various letter grades rated on their quality reviews. Eventually the DOE hoped to have enough data to determine whether the quality review was truly a predictor of future performance.

Table 5: Numbers of schools at each position of the consequences matrix, 2009

		Quality Review Rating				
		Outstanding	Well-Developed	Proficient	Underdeveloped w/ Proficient Features	Underdeveloped
Progress Report Grade	A	18	554	391	37	1
	B	0	105	116	13	2
	C	0	17	51	9	1
	D	0	1	17	3	5
	F	0	0	5	0	0

Note: Includes data from the 1346 schools that had both a progress report grade and a quality review rating in the 2008-2009 school year. All letter grades are from 2009; QR ratings are from 2008 and 2009, depending on the year of a school's last QR.
Source: Compiled from publicly available progress report results[30]

For instance, would the 18 C and D schools with "well-developed" QR ratings in Table 5 move into the A and B categories over the next few years? Would the three A and B schools with "underdeveloped" QR ratings receive a lower grade the following year? At the end of the 2009 school year, the tools were too new and had evolved too often to provide a valid link between the two measures.

Though the locus of accountability and the responsibility for the behaviors assessed in the QR process fell squarely on the school principal, the DOE theory of action eventually included the assumption that teams of teachers would use the various data tools to learn more deeply about students and respond more effectively to their needs. As the theory of action continued to evolve, Klein and the senior team began to talk about the strategy as having three phases. The first phase—consolidation—began in 2002 and was focused on the system. The second phase—the autonomy/accountability exchange—began in 2006, and was focused on schools. In 2009, the team began to talk about a third phase that was focused on classrooms. In order for accountability for student performance and team-based problem solving to become sustainable features of the culture at all levels of the system, they had to become part of the everyday work of teachers in classrooms.

TEACHER PERCEPTIONS OF ORGANIZATIONAL LEARNING AND ACCOUNTABILITY

Pushing the unit of analysis to the classroom level raises the question of how teachers perceive the accountability and organizational learning cultures in their schools, both of which are critical aspects of the DOE performance management system. This section draws on a large-scale survey of New York City teacher perceptions and provides an early look at the DOE's theory in action as reflected in teacher

perceptions of organizational learning and accountability at their schools. In particular, we explore the extent to which DOE schools do indeed fall out differently across these two dimensions of organizational learning and accountability, suggesting that differences in school cultures do exist and are worthy of further investigation. Second, we examine correlations between various indicators of performance and where schools fall on an organizational learning and accountability matrix. We stress that these are investigatory analyses; we offer several suggestions for future research at the end of this section to deepen our understanding of DOE's theory in action in schools.

The data we analyzed in this exploratory analysis comes from items we added to the DOE's 2009 Learning Environment Survey for teachers. This survey was launched in February of 2009 to approximately 80,000 teachers, of whom 58,000 teachers responded either on paper or online, reflecting a 73% response rate. Our data come from the 26,500 teachers who took the survey online, who make up 46% of the responding teachers. When we aggregate teacher responses to the school level in our analyses, we focus on non-special-education schools where at least 50% of teachers responded to a survey, and where at least half of these teachers responded to the survey online. This sample of majority-responding schools includes 19,500 teachers in 586 schools.

Building on a pilot validation study conducted in another large urban district, we investigated the extent to which teachers felt "psychologically safe" in their schools. This measure was adapted from Edmondson for the present context and is one of the core components of the organizational learning building block, "supportive work environment," previously discussed on page 11.[31] The "psychological safety" measure was composited from three items and added to the 2009 Learning Environment survey:

- In this school, it's easy to speak up about what is on your mind.

- People in this school are eager to share information about what does and doesn't work.

- People in this school are usually comfortable talking about problems and disagreements.

Teachers responded to these items on a 4-point response scale that ranged from *strongly agree* to *strongly disagree*. A principal components analysis (PCA) indicated that these three items are facets of a single construct.[32] We calculated the arithmetic mean of the items to produce a single value capturing each teacher's sense of psychological safety in their school.

The construct of "accountability" was composited from seven items on the teacher survey, four of which were existing measures on the Learning Environment Survey (first four below) and three of which were new items piloted in other schools prior to being added (remaining three). The items encapsulate many aspects of accountability covered in the literature by capturing teacher perceptions of both internal and external accountability, aspects of the accountability process, and outcomes that serve as accountability targets. Further, and consistent with O'Day's notion of the "new accountability," which places the school at the center of intervention and the teachers as the units for action, these items assessed teacher perceptions of their *school* (rather than their department or some other level of analysis):[33]

- My school has high expectations for all students.

- Teachers in this school set high standards for student work in their classes.

- This school makes it a priority to help students find the best ways to achieve their learning goals.

- My school has clear measures of progress for student achievement throughout the year.

- Our school is focused on improving performance on measures of student achievement for this year.

- Meeting targets for student progress is a priority in this school.

- Helping students reach targets for mastery of important skills and content is a priority for this school.

Conceptually, these items tap into teachers' perceptions of accountability for improving student learning. Together, they offer a new, robust representation of accountability because they cover both the processes and outcomes of accountability for achieving high-performance. This is consistent with many of the theories of action in education reform today, including NYC's theory of action, which has shifted away from a singular emphasis on outputs to one that emphasizes both inputs (e.g., Quality Reviews) and outputs (e.g., performance and progress reports).

Empirically, these items hold together as a single construct extremely well.[34] Therefore, as with the psychological safety items, we produced a single accountability value for each teacher by taking arithmetic means of the items.

Our statistical analyses also show that psychological safety and accountability are correlated, but this association is not surprising—nor does it negate the theoretical distinction between the constructs.[35] Both capture perceptions of school culture, which is why they are correlated, and yet they capture different aspects of that culture: psychological safety is about the conditions for adult learning, whereas accountability is about adults focusing on student learning. In fact, 140 schools in our sample were high on one measure but low on the other, suggesting that these two aspects of culture can exist independent of one another.

We aggregated the psychological safety and accountability measures across teachers at each school and reverse-coded them to produce school-wide values for psychological safety and accountability such that higher values reflect a higher sense of psychological safety and accountability in a school. We used the medians of these school-aggregate measures as the dividing line between "high" and "low" values on each measure. The schools that we refer to as "high" on psychological safety and accountability, therefore, are the 223 schools that had values at or above 3.08 and 3.42 on each measure, respectively. The 223 schools that had values below these cut-off points on both measures constitute our pool of schools that are "low" each measure. Figure 1 displays the variance that exists between schools regarding teacher perceptions of their school's ability to balance organizational learning and accountability.

Figure 1: School-level teacher responses regarding accountability/psychological safety (n=586)
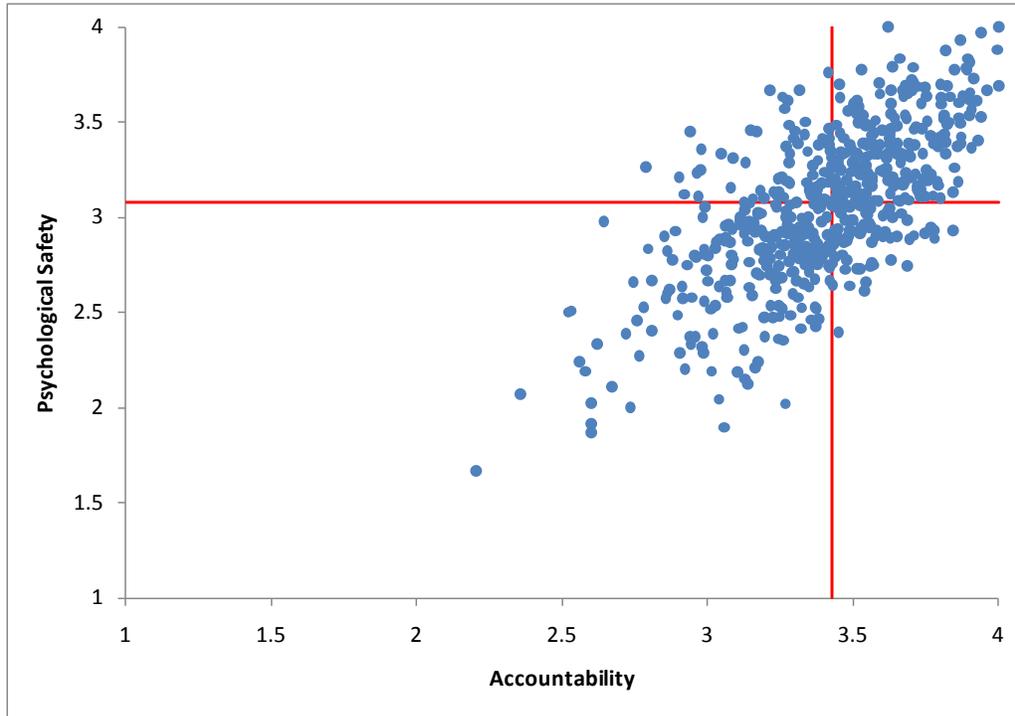
To investigate this further, we explored the off-diagonals of the matrix in Figure 1 to see if there were any apparent differences in performance between the two groups of schools, and we found that there were. As Table 6 below shows, schools with high accountability and high organizational learning environments outperform low accountability and low organizational learning environments along a number of performance indicators.[36] We again emphasize that these are preliminary analyses and note, for example, the difference in response rates across the categories. Still, these data do provide some early and suggestive evidence in support of the theory of action regarding the performance benefits associated with creating school cultures in which teachers feel both accountable and psychologically safe to invest in their own learning.

Table 6: Descriptive statistics for high psychological safety/high accountability and low psychological safety/low accountability schools

Performance Measure	Min.	Max.	High/High		Low/Low		t-statistic
			n	Mean(sd)	n	Mean (sd)	
2009 progress score	6.7	70	172	47.07 (9.1)	180	42.62 (10.4)	4.28***
2008 performance score	1.5	25	167	16.43 (.34)	177	14.05 (4.2)	5.14***
2009 performance score	7.2	25	172	19.65 (3.9)	180	17.64 (3.9)	4.86***
07-08 Quality Review	1	5	168	3.73 (.51)	189	3.44 (.64)	4.67***

* p<.05; ** p<.01; *** p<.001

CONCLUSION

The DOE's performance management system is a distinctive attempt to execute the high wire act of balancing support and pressure that scholars and practitioners have been writing and speaking about for the last decade. The DOE had an explicit theory of action for improving student outcomes—the autonomy/accountability exchange—and attempted to create organizational structures, systems, and culture that were coherent with this theory.[37] For the autonomy/accountability exchange to work at scale, schools had to develop the capacity for adult learning in order to create better learning outcomes for students, given that the central bureaucracy would no longer mandate particular curricular and instructional approaches. This is much more than a psychological contract between the central office and schools, however; expectations for performance are high and the consequences for not meeting the requirements of the accountability system are serious. Student performance in New York City, as measured by state exam scores and graduation rates, has accelerated since the evolving theory of action was put into practice, but the jury is still out on the specific contributors to that improvement. The state's dramatic upward adjustment to cut scores for 2010 created massive drops in the percentage of students reaching proficiency and reignited the debate about just how much progress students in New York City have actually made during Klein's tenure.

Nevertheless, the approach to performance management, balancing accountability and organizational learning, is already being adopted by other districts. This historical analysis and preliminary empirical work suggest several promising areas for future research. First, we suggest that scholars investigate the ways in which different districts have approached the tensions explored in this paper. Though the DOE's emphasis on the granting of autonomy in exchange for accountability regardless of schools' prior performance is being adopted in other districts, most cities still require schools to earn expanded autonomy. As the demands for improving performance continue and impatience with the status quo increases, it will be interesting to watch how different districts develop and enact their theories of change. How will these theories of action fare and can we create some generalizable principles about what kinds of interventions "work" to produce the results desired? Do different theories of action require different approaches to performance management? Are districts increasingly trying to balance accountability and organizational learning, and if so, do their attempts differ based on their particular theories of action? Investigations of these topics would add greatly to research and practice.

We can also ask, how do these systems impact school culture and what effect does this have on performance? Often, discussions of culture in the education sector are either amorphous ("schools need strong cultures") or descriptive (i.e., long lists of general characteristics such as "collaboration"). If we can identify and more clearly define specific dimensions of school culture (such as accountability and psychological safety) that "matter" in the daily practice of teachers and are predictive of improvements in learning outcomes of students, then we would be better positioned to find points of entry to effect change. If subsequent analyses do support the preliminary findings offered here, and we expect they will, then the next logical question to ask is "why?" Why is it that some schools are able to achieve a desirable balance between organizational learning factors such as psychological safety and a sense of accountability? Qualitative data are needed to answer this question and would complement the large-scale survey work that is currently underway in DOE and could provide insight into effects observed.

We hope that the chapters in this volume, including our description of and investigation into the DOE's performance management system and the theories that underlie it, will provide scholars and practitioners with a point of reference for considering their own efforts to cultivate high-performing systems of schools.

ENDNOTES

[1] We are especially grateful to the Public Education Leadership Project of Harvard Business School and Harvard Graduate School of Education for their support of this work and to Rebecca Holcombe for her research assistance. Portions of this paper are adapted from Stacey Childress and Tonika C. Clayton, *Focusing on Results at the New York City Department of Education* (Cambridge, MA: Public Education Leadership Project at Harvard University, PEL-054, Harvard Business Publishing, 2007). All direct quotes from Joel Klein and other NYCDOE staff are from this piece unless otherwise noted.

[2] Penny Bender Sebring and Anthony S. Bryk, "School Leadership and the Bottom Line in Chicago," *Phi Delta Kappan* 81, no. 6 (2000): 440.

[3] Quote from Arlene Ackerman in Stacey Childress and Jennifer M. Seusse, *Star Schools Initiative at the San Francisco Unified School District* (Cambridge, MA: Public Education Leadership Project at Harvard University, PEL-039, Harvard Business Publishing, 2006).

[4] Brian A. Jacob, "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics* 89, no. 5-6 (2005): 761–796.

[5] The description of the theory of action, strategy and structural changes in this section only cover information relevant to the performance management issues covered in this paper. Other papers in this volume offer more thorough discussions of these important topics.

[6] For an overview of the relevant work of Ouchi, Hill, and Barber see: William G. Ouchi and Lydia G. Segal, *Making Schools Work: A Revolutionary Plan to Get Your Children the Education they Need* (New York: Simon & Schuster, 2003); Paul Hill et al., *Portfolio School Districts for Big Cities: An Interim Report* (Seattle, Center on Reinventing Public Education, 2008), http://www.crpe.org/cs/crpe/download/csr_files/pub_psdp_interim_oct09.pdf; and Michael Barber, *Instruction to Deliver: Fighting to Transform Britain's Public Services* (London: Methuen, 2008).

[7] The number of schools grew to nearly 1,500 by 2009, fueled by new charter schools and the new small high schools that launched to replace large comprehensive high schools. Existing single high schools were often replaced by four to six smaller schools, contributing to the rapid growth in the absolute number of schools.

[8] Chrysan Gallucci, "Districtwide Instructional Reform: Using Sociocultural Theory to Link Professional Learning to Organizational Support," *American Journal of Education* 114, no. 4 (2008): 541–581.

[9] See, for example, J. Bradley Cousins, "Understanding Organizational Learning for Educational Leadership and School Reform," in *International Handbook of Educational Leadership and Administration*, ed. Kenneth Leithwood et al. (Dordrecht, Boston: Kluwer Academic, 1996), 589-652; and Kenneth Leithwood, Lawrence Leonard, and Lyn Sharratt, "Conditions Fostering Organizational Learning in Schools," *Educational Administration Quarterly* 34 (1998): 243-276.

[10] Barbara Levitt and James G. March, "Organizational Learning," *Annual Review of Sociology* 14, no. 1 (1988): 319–338.

[11] See, for example, Daniel A. Levinthal and James G. March, "The Myopia of Learning," *Strategic Management Journal* 14, (Winter, 1993): 95–112; and Barbara Levitt and James G. March, "Organizational Learning," *Annual Review of Sociology* 14, no. 1 (1988): 319–338.

[12] Meredith I. Honig, "District Central Offices as Learning Organizations: How Sociocultural and Organizational Learning Theories Elaborate District Central Office Administrators' Participation in Teaching and Learning Improvement Efforts," *American Journal of Education* 114, no. 4 (2008): 627–664.

[13] See for example, Lev S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, ed. Michael Cole (Cambridge: Harvard University Press, 1978); and James V. Wertsch, "A Sociocultural Approach to Socially Shared Cognition," in *Perspectives on Socially Shared Cognition,* eds. Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley (Hyattsville, MD: American Psychological Association, 1996): 85–100.

[14] For a review, see Michael S. Knapp, "How Can Organizational and Sociocultural Learning Theories Shed Light on District Instructional Reform?" *American Journal of Education* 114, no. 4 (2008): 521–539.

[15] For example, see Michael S. Knapp, "How Can Organizational and Sociocultural Learning Theories Shed Light on District Instructional Reform?" *American Journal of Education* 114, no. 4 (2008): 521-539; and Meredith I. Honig, "District Central Offices as Learning Organizations: How Sociocultural and Organizational Learning Theories Elaborate District Central Office Administrators' Participation in Teaching and Learning Improvement Efforts," *American Journal of Education* 114, no. 4 (2008): 627–664.

[16] David A. Garvin, Amy C. Edmondson, and Francesca Gino, "Is Yours a Learning Organization?" *Harvard Business Review* 86, no. 3 (2008): 109–116.

[17] Amy C. Edmondson, "Psychological Safety and Learning Behavior in Work Teams," *Administrative Science Quarterly* 44, no. 2 (1999): 350–383.

[18] David A. Garvin, Amy C. Edmondson, and Francesca Gino, "Is Yours a Learning Organization?" *Harvard Business Review* 86, no. 3 (2008): 109–116.

[19] Jennifer A. O'Day, "Complexity, Accountability, and School Improvement," *Harvard Educational Review* 72, no. 3 (2002): 293–392.

[20] Richard F. Elmore, *School Reform from the Inside Out: Policy, Practice, and Performance* (Cambridge, MA: Harvard Education Press, 2004).

[21] Email from Richard Elmore to author, June 3, 2010. For more Elmore on accountability, see "When Accountability Knocks, Will Anyone Answer?"and "Building a New Structure for School Leadership," both in Richard F. Elmore, *School Reform from the Inside Out: Policy, Practice, and Performance* (Cambridge, MA: Harvard Education Press, 2004).

[22] Amy C. Edmondson, "The Competitive Imperative of Learning." *Harvard Business Review* 86, no. 7 (Jul, 2008): 60–67.

[23] New York City Department of Education, "Progress Report Results for all Schools Citywide, 2007, 2008, and 2009," http://schools.nyc.gov/Accountability/tools/report/default.htm.

[24] For a more technical explanation of the calculation of the various subsections and overall progress report grade, see two documents on the NYCDOE website at this link: "Educator Guide: The New York City Progress Report - High School." The New York City Department of Education. http://schools.nyc.gov/NR/rdonlyres/DF48B29F-4672-4D16-BEEA-0C7E8FC5CBD5/70866/EducatorGuide_HS_1104092.pdf, 2010).

[25] New York City Department of Education, "Educator Guide: The New York City Progress Report, Elementary/Middle School." http://schools.nyc.gov/NR/rdonlyres/DF48B29F-4672-4D16-BEEA-0C7E8FC5CBD5/65798/EducatorGuide_EMS_0804094.pdf, 2010).

[26] Shael Polakow-Suransky, *In Defense of High School Progress Reports: Responding to Readers' Comments* (Gotham Schools, 2009), http://gothamschools.org/2009/11/25/responding-to-readers-comments/#more-28138

[27] Shael Suransky, conversation with authors, June 17, 2010.

[28] For a non-technical overview of Koretz's criticism, see his guest blog on Eduwonkette from September 17, 2008, *Guest Blogger Daniel Koretz on New York City's Progress Reports* (Eduwonkette: Education Week, 2008), http://blogs.edweek.org/edweek/eduwonkette/2008/09/guest_blogger_daniel_koretz_on_1.html

[29] New York City Independent Budget Office, "The School Accountability Initiative: Totaling the Cost" (2010), http://www.ibo.nyc.ny.us/iboreports/Schoolaccountability111308.pdf.

[30] New York Department of Education, "2008-09 Progress Report Results for all Schools Citywide," http://schools.nyc.gov/Accountability/tools/report/default.htm.

[31] Amy C. Edmondson, "The Competitive Imperative of Learning." *Harvard Business Review* 86, no. 7 (Jul, 2008): 60–67.

[32] The eigenvalue for the first component was 2.48, while the second component eigenvalue was well below the 1.0 cut-off point at 0.31. Cronbach's alpha for these three items was 0.89, indicating strong internal reliability. The items were not standardized or weighted, since standard deviations were similar across the three items (0.75 to 0.92), and PCA eigenvectors were nearly identical (0.57 to 0.59). This allowed us to preserve the original scale of this measure for ease of interpretation.

[33] Jennifer A. O'Day, "Complexity, Accountability, and School Improvement," Harvard Educational Review 72, no. 3 (2002): 293–392.

[34] Results from the PCA indicated that these seven items tap into a single construct, with the first component eigenvalue at 4.71, while the second component eigenvalue was 0.95. Cronbach's alpha showed a strong internal reliability, with a value of 0.92. As with the psychological safety items, PCA weights were not used (eigenvectors ranged from 0.33 to 0.40) and items were not standardized (standard deviations ranged from 0.63 to 0.77).

[35] Pearson's r for the psychological safety and accountability variables is 0.69.

[36] Notes on the performance measures employed in these preliminary analyses:

    a) We used the overall Progress Report Score (a grade of A-F) that each school receives each year.

    b) We used the 2007–08 Quality Review scores because only approximately 30% of the schools got 2008–09 Quality Reviews (if a school got an overall Progress Report grade of B or better and a "Proficient" or better on their 2007–08 quality review, they were not reviewed in 2008–09). The Quality Review scores were based on a 1 to 5 scale, which we reverse coded so that 5 = excellent and 1 = underprepared (of which there was only 1 school in our sample).

    c) Both the progress and performance scores are complex calculations made by the district; we have included the min/max to the side in case this is helpful for contextualizing.

[37] Stacey Childress, Richard Elmore, Allen Grossman, and Caroline King, *Note on the PELP Coherence Framework* (Cambridge, MA: Public Education Leadership Project at Harvard University, PEL-010, Harvard Business Publishing, revised 2007), http://cb.hbsp.harvard.edu/cb/web/product_detail.seam?R=PEL010-PDF-ENG&conversationId=143801&E=70507.