The Paradoxes of 
Data-Driven School Reform

Learning from Two Generations of Centralized 
Accountability Systems in the United States

HEINRICH MINTROP
GAIL L. SUNDERMAN

In organizations, the use of data is interwoven with forms of organizational management, and data infrastructures are embedded in management structures. With the advent of the accountability era of the 1990s in the United States, a new form of organizational management took hold in the public school sector that uses data to concentrate goal setting and performance monitoring functions at the state and, later, federal levels. Sennett analyzes this new centralization as a confluence of two new developments in the organization of work, one technical, the other political: (1) new data warehousing technologies and infrastructures made it possible, with relatively little administrative capacity, to set targets based on a small set of quantitative performance indicators and monitor whether these targets were reached by large numbers of relatively small performance units; and (2) increasingly deregulated workplaces, spreading from private industry into the public sector, suggested to political and managerial elites to link performance indicators and data to sanctions imposed on employees for lack of performance.1

In education, this has given rise to accountability systems driven by performance on standardized achievement tests. Over the last twenty years or so, accountability systems have evolved in what we see as three distinct design generations up to this point. The first generation began with a handful of pioneering states that each developed its own system designs in the
early to mid-1990s. A second-generation pattern took over when the No Child Left Behind Act of 2001 (NCLB) made sanctions-based accountability universal across all federal states. At this juncture, ten years after the passage of NCLB, with many states opting out of its main requirements, policymakers and school systems are relying on a new set of design components that suggest the beginning of a third generation, this one pivoting on educator evaluations.

Generations have a common ancestry. And so it is with school accountability in the United States. All three generations of accountability designs share common elements—incentives, measureable goals, performance monitoring and assessment, and consequences for low performance. All three give primacy to central administrative levers and managerial leadership to guide the system and penetrate the technical core. In so doing, they reflect trends in public management that have been discussed in the public administration literature under the auspices of "reinventing government" or "New Public Management." But these structural commonalities play out quite differently in each generation, as old system elements are abandoned and new ones are added. As each generation follows the next, we need to learn from the experiences of the previous one to avoid repeating mistakes.

We proceed in this chapter in four steps. We analyze the two successive generations in the first two sections, focusing on system functionality and outcomes. As to functionality, we look at the match between managerial goals and required capacities. Dysfunctionalities arise when goal-capacity discrepancies undermine incentives. As to student learning, we look at its impact on test scores, curriculum, and instruction. In a third step, we discuss possible theoretical explanations for the patterns we identify. In a final step, we apply these explanations to implementation scenarios for third-generation designs.

Our argument is that high-stakes accountability in the first two generations worked and did not work. The systems have turned out to be powerful motivators for educators to reach the state-sponsored goals while, overall, missing their substantive educational goals of increasing the quality of students’ learning experience and closing the achievement gap. At the same time, the systems produced dysfunctionalities that strained the problem-solving capacity at each level. Drawing on public administration literature, we suggest that these patterns might be explained by a set of paradoxes: politicized goal setting, unintended technical inflexibilities, unforeseen capacities and capacity deficits, indicator corruption, and performativity. These paradoxes are loosely understood as developments or outcomes that are contrary to beliefs about how accountability should work. To mitigate these five paradoxes for the third-generation design, we argue for a more comprehensive institutional framework that lessens the primacy of administrative levers in favor of mobilizing teachers’ concerns around good instruction.

THE FIRST GENERATION

Given that states were key players in the early designs of the 1990s, system designs differed according to state preferences. Some states, such as Maryland and Kentucky, fused the demand for immediate results with a quest for aspirational educational goals that were given teeth in performance-based tasks on state assessments. Other states, notably Texas, chose a lower cognitive goal horizon. In short order it became clear that these systems functioned quite differently. A state system such as Maryland’s, with very ambitious cognitive and pedagogical goals combined with low levels of state support for school improvement, could not sustain the momentum needed to post continuous growth on state assessments and had to retreat on its goals to ensure system survival. A state system such as Texas’, with less ambitious goals, posted impressive growth on state tests, largely in the basic skills dimension, and produced fewer failing schools. These different trajectories are illustrated in figure 1.1.

First-generation systems showed that the functionality of the systems depended on the interplay among the difficulty of goals (i.e., realism of growth targets, cognitive rigor of state tests), available capacity reserves, intensity of external pressure, and educators’ internal valuation of accountability goals. Incentive systems, by design, locate performance problems in lack of employee motivation, here presumed lack of teacher effort and low expectations. The primary accountability unit was therefore the school, which, incentivized by information, clear goals, rewards and sanctions, was expected to search for new ways of improving its collective performance. Middle management (i.e., districts), presumably unwieldy, rule bound, and bureaucratic, were left out of some of the early state designs. Schools would deliver immediate and continuous quantitative results toward envisioned performance targets. Organizational learning and motivational potential, heretofore untapped, would make costly investments in capacity building unnecessary.

The systems were indeed powerful. Growth targets, performance assessments, and the threat of sanctions compelled schools and districts to pay close attention to system demands. But the systems' motivational power
was double edged. The literature noted an upsurge of energy elicited by the threat of sanctions, sometimes leading to impressive turn-around.\(^8\) The upsurge could just as well turn to demoralization, however, when test score gains were not forthcoming or could not be sustained, and when perceived unmanageable threats led to organizational rigidity.\(^9\) Demoralization was exacerbated as sanctions were applied to whole organizations, hence to high- and low-performing teachers and principals in equal measure.\(^10\) While teachers embraced accountability in general, reaching closely to the test, though probably widely practiced, was seen as a violation of professional norms of good educational practice.\(^11\)

While in most pioneering states test scores on high-stakes assessments rose, it became apparent fairly early on that the number of schools facing sanctions was quite high, and a good portion of schools required external support to make it through the system. States, however, were reluctant to build up the required support and intervention capacity. Instead, they adjusted system designs to existing capacities.\(^12\) Toward the end of this phase, districts were increasingly brought back into the picture—for example, in the 1999 California design—as key missing links for capacity building but also as prime recipients of accountability sanctions.\(^13\)

In sum, it became apparent during this phase that performance information plus incentives eliciting educators’ efforts would not be sufficiently powerful without more extensive state or district support and an intervention infrastructure. High educational aspirations clashed with the specificity of goals that needed to be translated into immediate results. On the face of it, the Texas system, designed with modest cognitive goals attached to simple metrics, seemed viable. Here, high pressure was placed on increasing the number of students reaching a specific performance threshold that was within reach of most schools.\(^14\) The system looked promising as test scores kept rising without the need for expensive new investments in capacity building.\(^15\) Thus, new quantitative metrics coupled with new central data warehousing capabilities and added incentives enabled state governments to powerfully influence educators’ work. But problems with capacity and with steering the system toward more complex learning outcomes surfaced.

THE SECOND GENERATION

With the signing of NCLB in 2001, high-stakes accountability was made universal across all fifty states. Even though the experience of first-generation systems could have warranted more caution, the NCLB design greatly enlarged ambitions and further concentrated central control. The negative repercussions visible in first-generation experiments—namely, the difficulty for high aspirations toward complex learning to survive in a data-driven, “results now” environment and the failure of states to develop support and intervention programs commensurate to the frequency of school failure—was not imputed. Instead, with NCLB, high-stakes accountability became attached to an even purer incentive and sanctions regime stretched over an even larger distance between a goal setting center (i.e., federal level) and implementing agencies (i.e., schools and districts) than had heretofore been the case.

The quantitative performance measure became at once extremely simple (i.e., the percent of students scoring proficient on state achievement tests) and more fine-grained (i.e., separate subgroup goals for each statistically significant student subpopulation). Districts were now included as key accountability recipients. The timeline became fixed and shortened (i.e., proficiency by 2013–2014), and growth quotas toward the end goal were expected to be met. The sanctions regime was to unfold in preordained stages of increasingly severe interventions and sanctions, not adjustable to states’ available support capacity. With the control of performance measures and sanctions
moving to the federal level and state governments retaining control over assessments, proficiency definitions, and investments in support and intervention, the distance between goal setting and substantive implementation stretched more widely.

Thus, NCLB followed from assumptions not dissimilar to first-generation rationales, but with inflated expectations. Even more so than the first-generation systems, the NCLB sanctions regime was based on the assumption that recalcitrant local actors often block reform and that imposing reform by a distant force less entangled in local pressures, bureaucratic institutions, and politics will be more successful. It therefore needed to bank, to a considerably large extent, on the presumed power of high-stakes incentives to penetrate the system to the core and on the willingness of the states (unsanctioned by the federal government in this respect) to do the right thing: maintain ambitious educational goals and make the necessary investments to ameliorate deficits where the power of sanctions proved to be insufficient.

Measured by high-stakes state tests, the system was a success: test scores in most states tended to rise continuously. As figure 1.2 indicates, these gains could not be confirmed with gains on low-stakes tests, namely the National Assessment of Educational Progress (NAEP). NAEP scores demonstrated that the achievement gap, certain modest improvements in achievement notwithstanding, lingered throughout the NCLB years. NCLB failed to accelerate the rate of achievement growth or to meet its overarching goal of proficiency for all by 2014 by a long shot. Given the discrepancy between high-stakes and low-stakes test results, it is not clear what the high-stakes state tests were actually measuring. To some, a good dose of teaching to the high-stakes test explained some of the gains.

Evidence accumulated that state governments under NCLB, not unlike the more ambitious state systems of the first generation, reduced the rigor of their assessments over time to keep the systems functional and intervention burden manageable. Lacking capacity to intervene, states relied on districts, under tremendous pressure themselves, to support schools that persistently missed their performance targets. Struggling with their own limited administrative capacity and professional competence, many districts seem to have responded as a low-capacity bureaucracy would, with relatively low-cost simplistic strategies. Test alignment, remediation, benchmarking, consultants who helped statistically identify “bubble kids,” prescriptive and packaged instructional programs such as literacy and remediation programs, excessive practice and reteaching, and the like helped reinforce a low-rigor learning culture in schools.

![Figure 1.2](image)

**Discrepancy between high-stakes and low-stakes performance, 2002–2005**

*Note: Grade math proficiency trend based on state assessment v. NAEP; N=25 states. Source: Lee Jaeyung. Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-Depth Look into National and State Reading and Math Outcome Trends (Cambridge, MA: The Civil Rights Project at Harvard University, 2006).*

The design worked as intended, and it didn’t. It is probably safe to say that, during its ten-year run, NCLB became the decisive impetus for an intense dynamism of educational reform. Specifically related to the NCLB design, performance deficits of previously marginalized groups were taken more seriously due to the sting of subgroup-based sanctions. Yet, the system produced dysfunctions that strained the problem-solving capability at each level. Evidence has accumulated that federal goal setting created much movement around the proficiency cut-off points but was unable to register growth in the below-proficiency band where many high-poverty schools operate, was insensitive to the exclusion of students, created an undue burden for schools with highly heterogeneous student populations with many subgroups, and resulted in an untenable number of failing and misidentified schools in many state systems.

In particular, sanctions connected to student groups whose learning needs are not easily standardized, such as special education students or English language learners, converted many otherwise solidly performing schools into failing ones. In the logic of the NCLB design, these groups, when excluded from the sanctions regime, would be pushed to the sidelines, neglected by
schools under high pressure; but when included, they would make growth goals quite daunting, undermining the incentive function.

Inflexibility of growth targets and sanctions stages all but doomed the functionality of state systems that held on to rigorous standards and also faced challenging student demographics. The signaling function of the incentive system lost its punch as large numbers of schools and districts entered the corrective action and sanctions stages. In California, for example, more than half of all schools and districts had done so after seven years of NCLB. High numbers of failing schools and districts linked to an increasing severity of corrective action strategies dissipated states' limited intervention capacities.

Especially in school districts serving disadvantaged populations, high-stakes pressure encouraged job design structures that were apt to reward teachers with high compliance dispositions and repulse educators with a high need for personal growth. Studies suggest that rather than freeing up operational units for innovation and learning, high pressure coupled with simple metrics and the demand for scientifically proven interventions resulted in a degree of micromanagement not seen under the most rule-bound traditional bureaucracy. This may have been justifiable if the systems had produced large and confirmable equity gains in the low-rigor basic skills dimension. But this has not been the case.

In view of these patterns, the basic architecture of NCLB seems to have failed in helping us reach the law’s most important goals. Despite enormous district and school efforts, as documented by sizable increases in state test scores, second-generation accountability designs were dysfunctional in many states and failed to produce substantial improvements in schools’ learning cultures and confirmed student outcomes. With the fading of NCLB, we need to part with the idea that a bare-bones growth quota and sanctions regime with an artificial end goal far above the stretch capacities of the system and far removed from the political realities and educational complexities of implementation in states, districts, and schools could be powerful enough to close the achievement gap and produce a high-quality learning culture in short order.

THEORETICAL EXPLANATIONS

The purpose of this chapter is not to explain why the NCLB policy regime may have failed in light of its own goal, at its most far-reaching extreme, of closing the achievement gap within a historically unprecedented thir-teen years. There is scholarship on the power of extraschool factors, most notably socioeconomic factors, that might provide some useful explanations that, in our eyes, have clearly gained more credibility as a result of the failed NCLB experiment. Our purpose here is narrower. We need to understand how a centralized public management system with commendable social goals that holds employees accountable to outcomes with standardized data could fail a much more modest quest—that is, to put schools and districts on a continuous path of improvement in key dimensions of proper functioning. In other words, we limit our discussion to the managerial aspects of the problem. Drawing from the literature on public management, we try out a number of explanations that may help us understand the phenomena we identified earlier.

High-stakes data-driven accountability in the educational sector is the education-specific version of what is referred to in the literature as the new public management (NPM). Borrowing from private-sector principles, its founding doctrine advocated a move away from bureaucratic rule-bound service provisions to entrepreneurial management based on establishing explicit, quantitative performance indicators, incentivizing outputs, separating steering by output controls from operations, bringing accountability to smaller production units, and, in the process, cutting costs, increasing labor discipline, and stimulating organizational innovation.

As NPM systems all over the world have aged, the literature on this phenomenon has matured as well. It is now widely acknowledged, even by its theoretical protagonists, that, as with other management models that came before it, initial claims for NPM models proved to be inflated, and the idea that a new organizational device could solve entrenched social problems, preferably without the outlay of new resources, looks ever more ideological than rational. New paradoxes, unanticipated developments, and unintended effects have been detected. Five of those seem especially applicable to high-stakes accountability systems in education: politicized goal setting, technical inflexibility, capacity deficits, indicator corruption, and performativity. Understanding these patterns can help inform the design of the next generation of policies.

Politicized Goal Setting

The linchpin of NPM designs—ambitious goal setting by the center, flattened bureaucracies in the middle, freeing work from rule-bound regulation, and learning and problem solving on the bottom—were compromised once they encountered the entrenched political and bureaucratic structures
of educational systems. In contrast to the presumed rationality of private-sector functioning, educational goals formulated by policymakers were still subject to politicization; and while middle layers at the regional or local levels may have lost some control, they did not vanish. Under these circumstances, the result was a tendency to depoliticize and rebuild the system. In the United States, high-stakes accountability was accompanied by highly politicized goals formulated with the customary bravado of the policy-making center (e.g., being number one in math and science in the year 2000; closing the achievement gap in the year 2014) that were wholly unrealistic by any standards of rational management and, fixed in policy and caught in political posturing, could only be revised with great difficulty.

When first-generation systems were designed, districts were left out of the picture. In some pioneering states (e.g., Maryland and Kentucky), they weren’t direct recipients of sanctions, and neither were the structures of local governance or administration changed. Especially after the passage of NCLB, middle managers in districts were not made directly accountable, but they responded vigorously with increasing regulation and micro-management of instruction, an approach that may make sense in light of customary bureaucratic practices, prevailing learning theories, and limited instructional know-how. Such a dynamic would tend to incentivize quiet acceptance of external regulation on the part of workers, rather than lively problem solving, producing increased conformance and isomorphism while diminishing aspirations outside of the narrow bounds of prescriptions, benchmarks, tests, and sanctions.

**Technical Inflexibility**

First-generation systems started out with very ambitious goals as well, but they were softened or rolled back as high performance demands ran up against high failure rates. Second-generation systems, by contrast, became victims of their own political rhetoric. While there was considerable push back by states and local education agencies to many of the NCLB accountability requirements, the basic framework remained unchanged. At the core, two-way communication between top and bottom was weakened by the adversarial stance of a demanding center that suspected “excuses” whenever performance did not rise up to expectations. The result was that the top (the federal government under NCLB) could not learn from the feedback that lower levels of the system (states, districts, and schools) could provide, creating what Collingridge calls an “inflexible technology” that forestalls “normal trial-and-error learning” within the system. Data may flow up and down infrastructural lines, but they do not speak.

**Unforeseen Capacities and Capacity Deficits**

Capacity building played an insignificant role in the theorizing about new outcome-driven public management systems, since these systems were designed with the arguable idea that lack of motivation on the part of employees is responsible for performance deficits and that powerful incentive systems can remedy this situation. We know now that this “willfulness” hypothesis is right and wrong at the same time. The determined focus of schools to improve scores on high-stakes tests may speak to the motivational power of incentives, and the continuous increase in high-stakes test results without concomitant verification by low-stakes tests speaks to a very specific capacity of schools and districts to streamline their operations to correspond to the measured performance indicators.

But there is a set of persistently low-performing schools, often located in impoverished community environments, that does not even have the capacity to do that due to their lack of basic teaching skills and organizational instability. The literature has pointed to the lack of effective support and intervention programs for this set of schools. Episodic infusions of discretionary money or changing modes of governance (e.g., charter convergence) have not produced consistent results. Capacity building for this set of schools may need to go beyond the managerial dimension (e.g., data infrastructure, outcome monitoring, work incentives, episodic resource allocation, governance changes). More comprehensive remedies would involve multiple institutions beyond the public school system that aim to strengthen the supply of qualified teachers and administrators, improve school-community relationships, and target extraschool risk conditions.

**Indicator Corruption**

At first glance, new public management systems, as well as high-stakes accountability, stress design characteristics such as clarity, consistency, authoritativeness, and stability that would seem to bode well for effective implementation. After all, the systems are presumably clear and specific in their simplicity, accurate in their measurability, authoritative in their top-down execution, and stable in their relentlessness, relative to far more loosely structured policy initiatives of the past. And indeed this upturn in rationality may explain their penetrating power.
But clarity and simplicity in design are vulnerable to indicator corruption, a phenomenon widely recognized in a variety of literatures. In the public administration literature, Campbell’s dictum about large-scale evaluations of planned change is famous: “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.”^41^ Holmstrom and Milgrom, drawing from principal-agent theoretical models, hypothesize that employers, when incentivizing complex work with rewards based on simple performance indicators, will encourage distortions as employees tend to concentrate efforts on only those dimensions for which they will be rewarded.^44^ NPM theorists have pointed to work distortions and indicator inflation in outcome-driven management and accountability systems in many countries and various areas of public administration and government.^45^ It is perhaps no surprise that these same distortions were found in educational accountability systems.

**Performativity**

Indicator corruption might be useful in explaining creative compliance or indicator inflation, such as teaching to the test or concentrating on “bubble kids,” but it does not seem powerful enough to explain the paradoxical disconnect between output and outcome that NPM systems seem to produce.^46^ The most visible sign of this disconnect in present-day U.S. education is the discrepancy between high-stakes and low-stakes test results. That is, while the systems produce the demanded output, they fail to produce the outcomes on which this output is supposed to deliver, such as substantive learning gains that can travel from school to life (or at least from one test to another) and young people’s engagement in their learning (to at least make it past high school).

Ball, drawing from the work of Lyotard, has attempted to capture this pattern of deep seriousness about outputs divorced from valued outcomes as performativity.^47^ In simplified terms, performativity in Lyotard’s formulation describes the tendency of postmodern society to legitimate knowledge, truth, and right action in terms of “what works” as measured by increasingly more fine-grained scientific measures, objective indicators, and defined externalized performances that apply to narrow roles or organizational fields.^48^ Performativity, thus, freezes out the “bigger picture” and ethical judgments that were previously interpreted by, and derived from, grand narratives, such as religion or prominent ideologies (e.g., liberalism, Marxism). When we become preoccupied with what works, we give up on what we ultimately value—our aspirations. Output trumps outcome. The performances themselves “stand for, encapsulate or represent the worth, quality or value of an individual or organization within a field of judgment.”^49^ Ball uses the concept of performativity to describe what occurs when a proliferation of externally imposed measures, indicators, rewards, sanctions, programmatic prescriptions that in the present day carry the aura of authority and scientific objectivity become internalized as appraisals of self and others. In the performativity pattern, the traditional defensiveness against bureaucratic intrusion and inwardness of teachers’ communications with learners in the enclosed space of classrooms turns into defensiveness and externalized auditability and “fabrication.” Fabricated practices, Ball stresses, are not mere artifice but are acted out for real, sometimes with earnestness and conviction.^50^ Yet, enacting them may still invoke personal struggle and a sense of lost authenticity and judgment. The performativity pattern may explain how accountability systems of both generations could succeed in producing output while neglecting to produce valued outcomes.

**The Paradoxes Put Together**

Of the five paradoxes of new public management regimes discussed here, two of them—capacity deficits and indicator corruption—were already well recognized in the analysis of earlier modes of public administration and management. But under NPM they appear in a different light. Capacity appears double edged. There is the capacity to fabricate audited performance, to speak with Ball, resulting in demanded output, and there is the capacity to produce valued outcomes. The imposition of NPM-like systems may have revealed astounding proportions of “fabrication capacity” in schools and districts, while the capacity to produce valued outcomes may have actually declined, most notably due to simple inattention.

Indicator corruption may have to be rethought in the era of NPM. As part of a much more comprehensive work regime and a concomitant society-wide social psychology, indicator corruption may fold into the more thorough performativity pattern. The paradox of technical inflexibility is unique to NPM. It is indeed paradoxical that a management mode explicitly designed to free schools from bureaucratic rules and regulations in order to stimulate problem solving close to clients would end up micromanaging schools with ever more tightly interwoven outcome and process controls that rigidified to the degree that the politicized top “management” of the system could not learn from its faulty goal setting.
In light of the experiences of the two preceding accountability generations and the analytical literature on modes of NPM, we need to design third-generation systems that avoid the traps of politicized goal setting, technical inflexibility, capacity deficits, indicator corruption, and performativity lest we continue the disconnect of output and outcome.

THIRD GENERATION

With the Common Core State Standards project, the federal Race to the Top initiative, the School Improvement Grants program, and the ESEA flexibility waiver program, high-stakes accountability has entered a new design phase, even though it is not clear if this loosely patched package of initiatives will survive state-level or congressional politics.\textsuperscript{31} Even so, new accents are set:

- The Common Core State Standards reintroduce aspirational educational goals, namely college readiness.
- The simplistic sanctions regime is abandoned, at least at the federal level and, with it, realistically, all inflated hope that management by results via quantitative growth quota disaggregated to the level of subgroups could close the achievement gap in short order.
- Investment in support and intervention is concentrated in the bottom 5 percent of the most needy schools.
- The demand to evaluate educators based on achievement tests and measures of practice attenuates the centralization of goal setting by sharing the control over data between center and local districts and schools, while at the same time requirements for extended data systems will enable "management" to measure performance with finer grain size and down to ever smaller performance units.

We assume for the moment that these are indeed the contours of third-generation designs to come, especially the combination of unsanctioned aspirational educational system goals and sanctioned individual teacher performance. Inferring from the experience of the previous two generations, we have a new system that can potentially go in either of two directions: greatly expand on the educational experience of students by stimulating instructional quality and student learning gains in tandem, or greatly reduce it by extending the reach of rebureaucratized levers of instructional micromanagement into the last vestiges of teachers' classroom work. The direction, we infer from previous generations, depends on the way the management system becomes embedded into the larger institutional infrastructure.

There can be little doubt that a high-stakes system that drills down to the level of the individual teacher will become extraordinarily powerful in motivating work behavior. But what type of modal behavior will be motivated is not clear. One scenario imagines a highly problematic repeat of the patterns from previous generations. To begin with, there is presently an intense debate on the reliability and validity of value-added measures calculated for individual teachers.\textsuperscript{32} For all practical purposes, the motivational power of these scores will be decided by the recipients, regardless of statistical properties, based on the scores' practical correspondence with a host of other quality criteria held by teachers. If the value-added scores repeatedly appear arbitrary, they may be discounted as valid judgments of work, but this does not mean that they are inconsequential. If the performativity pattern holds, they could be at once discounted and accepted as objective truth, in the face of which individual teachers would defensively make themselves "auditable" and produce the demanded output given their fabrication capacity.

Third-generation designs move key improvement levers from outcomes back to inputs by standardizing the evaluation of instruction and attaching sanctions to these evaluations for ineffective teachers. Presumably, a paramount purpose of this procedure is upgrading teaching quality. This in turn requires complex tools for judging the planning and execution of lessons that can structure deep analytical conversations between evaluating and evaluated parties. It is likely that this kind of sophistication is a stretch for the instructional leadership capacity of many local schools and districts, especially those at the lower end of the performance distribution. The default reaction to this situation of high-pressure, low-capacity, but high-"scientific" legitimacy is, as we have seen, a breaking down of task complexity and a simplified version of external (in this case district) performance monitoring per checklist. In this scenario, evaluations may become trivialized and discounted by high-performing teachers.\textsuperscript{53} An upgrading of instructional quality becomes stalled, and the negative dynamics of previous generations become more likely. Performativity is heightened when high test pressures and trivialized teacher evaluations go hand in hand.

In a more sanguine scenario, the higher-performing segment of teaching faculties in schools and districts embrace the performance assessments. The assessments turn out to generate robust judgments of teaching quality that are consistent with classroom observations and other quality criteria that
practitioners may intuitively entertain. That is, high-performing teachers feel validated in their expertise and effort. Districts eschew their tendency to attach simple monitoring devices to high-pressure situations and share responsibility for evaluations with expert educators, though the latter may be relatively few in number initially. As a result, judging instructional quality is a matter of measurement and instructional expertise. Districts oversee the process, external experts support it, and internal experts are empowered to set the norms of excellence for faculties.

Good instruction is deeply embedded in a web of interconnected and highly institutionalized relationships and links: community relations, professional connections, administrative hierarchies, and societal standards of proper comportment and required skills and knowledge. Community relations exert influence on teachers' work when local elites or parents, oftentimes in direct personal contact, voice their preferences and interact with local officials and educators. As professional workers, educators draw from a fund of established knowledge and internalized norms of good education and acceptable performance. They receive professional socialization and special training in higher education institutions. Internal relationships with colleagues in schools and interactions with external experts maintain and reinforce professional standards, norms, and skills. Administrative means of influencing instruction (i.e., rules, resource allocations, rewards, and sanctions for teachers and students, most notably performance, career, and remuneration criteria) reinforce these professional standards.

Third-generation designs that avoid preceding patterns need to avoid inflexible accountability "technologies." They need to attenuate politicized goals that are currently disconnected from the real growth potential of the system. Goals need to be challenging and persistent, but high pressure and relentlessness fosters rigidities that squelch aspirations and learning. Feedback loops that help goal-setting executives learn from experiences and adjust their expectations are essential. To turn a phrase from Osborne and Gaebler: those who row the boat cannot but help to watch the ones that steer; but those who steer need to train their eyes on the crew as they decide what course to chart and how far to go. To counterweigh the tendency of district administrations to respond to performance pressures with micro-controls of teaching, instructional experts—partly housed within district administrations, partly housed in the professional sphere outside of the K–12 school system—need to become key actors in teacher evaluations.

Given levels of capacity, embedding new complex standards for teaching into daily practice will require broad development of new skills and competencies among teachers and administrators and systematic recruitment of excelling instructional leaders. This is a task that surpasses the capacity of management and administrations. It brings professional organizations of all kinds—universities, associations, reform organizations, and new institutional bodies—into the center of the picture. An important step has been taken in shifting key accountability functions downward, back to local systems. It is imperative to recognize that instructional excellence probably cannot emanate from the perch of the center or the primacy of administration; it must be an internal striving of actively participating professional workers who are sensitive to articulated community needs.