Data–Driven Improvement and Accountability

Andy Hargreaves and Henry Braun

Boston College

October 2013

National Education Policy Center

School of Education, University of Colorado Boulder
Boulder, CO 80309-0249
Telephone: (802) 383-0058

Email: NEPC@colorado.edu
http://nepc.colorado.edu

This is one of a series of briefs made possible in part by funding from
The Great Lakes Center for Education Research and Practice and the Ford Foundation.

http://www.greatlakescenter.org
GreatLakesCenter@greatlakescenter.org

Suggested Citation:
Hargreaves, A. & Braun, H. (2013). *Data-Driven Improvement and Accountability.* Boulder, CO: National Education Policy Center. Retrieved [date] from http://nepc.colorado.edu/publication/data-driven-improvement-accountability/.

Data-Driven Improvement and Accountability

Andy Hargreaves and Henry Braun, Boston College

Executive Summary

The drive to enhance learning outcomes has assumed increasing salience over the last three decades. These outcomes include both high levels of tested achievement for all students and eliminating gaps in achievement among different sub-populations (raising the bar and closing the gap). This policy brief examines policies and practices concerning the use of data to inform school improvement strategies and to provide information for accountability. We term this twin-pronged movement, data-driven improvement and accountability (DDIA).

Although educational accountability is meant to contribute to improvement, there are often tensions and sometimes direct conflicts between the twin purposes of improvement and accountability. These are most likely to be resolved when there is collaborative involvement in data collection and analysis, collective responsibility for improvement, and a consensus that the indicators and metrics involved in DDIA are accurate, meaningful, fair, broad and balanced. When these conditions are absent, improvement efforts and outcomes-based accountability can work at cross-purposes, resulting in distraction from core purposes, gaming of the system and even outright corruption and cheating. This is particularly the case when test-based accountability mandates punitive consequences for failing to meet numerical targets that have been determined arbitrarily and imposed hierarchically.

Data that are timely and useful in terms of providing feedback that enables teachers, schools and systems to act and intervene to raise performance or remedy problems are essential to enhancing teaching effectiveness and to addressing systemic improvement at all levels. At the same time, the demands of public accountability require transparency with respect to operations and outcomes, and this calls for data that are relevant, accurate and accessible to public interpretation. Data that are not relevant skew the focus of accountability. Data that are inaccurate undermine the credibility of accountability. And data that are incomprehensible betray the intent of public accountability. Good data and good practices of data use not only are essential to ensuring improvement in the face of accountability, but also are integral to the pursuit of constructive accountability.

Data-driven improvement and accountability can lead either to greater quality, equity and integrity, or to deterioration of services and distraction from core purposes. The question addressed by this brief is what factors and forces can lead DDIA to generate more positive and fewer negative outcomes in relation to both improvement and accountability.

The challenge of productively combining improvement and accountability is not confined to public education. It arises in many other sectors too. This brief reviews evidence and provides illustrative examples of data use in business and sports in order to compare practices in these sectors with data use in public education. The brief discusses research and findings related to DDIA in education within and beyond the United States, and makes particular reference to our own recent study of a system-wide educational reform strategy in the province of Ontario, Canada.

Drawing on these reviews of existing research and illustrative examples across sectors, the brief then examines five key factors that influence the success or failure of DDIA systems in public education:

1. The nature and scope of the data employed by the improvement and accountability systems, as well as the relationships and interactions among them;
2. The types of indicators (summary statistics) used to track progress or to make comparisons among schools and districts;
3. The interactions between the improvement and accountability systems;
4. The kinds of consequences attached to high and low performance and how those consequences are distributed;
5. The culture and context of data use -- the ways in which data are collected, interpreted and acted upon by communities of educators, as well as by those who direct or regulate their work.

In general, we find that over more than two decades, through accumulating statewide initiatives in DDIA and then in the successive Federal initiatives of the No Child Left Behind Act and Race to the Top, DDIA in the U.S. has come to exert increasingly adverse effects on public education, because high-stakes and high-threat accountability, rather than improvement alone, or improvement and accountability together, have become the prime drivers of educational change. This, in turn, has exerted adverse and perverse effects on attempts to secure improvement in educational quality and equity. The result is that, in the U.S., Data-Driven Improvement and Accountability has often turned out to be Data Driven Accountability at the cost of authentic and sustainable improvement.

Contrary to the practices of countries with high performance on international assessments, and of high performing organizations in business and sports, DDIA in the U.S. has been skewed towards accountability over improvement. Targets, indicators, and metrics have been narrow rather than broad, inaccurately defined and problematically applied. Test score data have been collected and reported over too short timescales that make them unreliable for purposes of accountability, or reported long after the student populations to which they apply have moved on, so that they have little or no direct value for improvement purposes. DDIA in the U.S. has focused on what is easily measured rather than on what is educationally valued. It holds schools and districts accountability for effective delivery of results, but without holding system leaders accountable for providing the resources and conditions that are necessary to secure those results.

In the U.S., the high-stakes, high-pressure environment of educational accountability, in which arbitrary numerical targets are hierarchically imposed, has led to extensive gaming and continuing disruptions of the system, with unacceptable consequences for the learning and achievement of the most disadvantaged students. These perverse consequences include loss of learning time by repeatedly teaching to the test; narrowing of the curriculum to that which is easily tested; concentrating undue attention on "bubble" students near the threshold target of required achievement at the expense of high-needs students whose current performance falls further below the threshold; constant rotation of principals and teachers in and out of schools where students' lives already have high instability; and criminally culpable cheating.

Lastly, when accountability is prioritized over improvement, DDIA neither helps educators make better pedagogical judgments nor enhances educators' knowledge of and relationships with their students. Instead of being informed by the evidence, educators become driven to distraction by narrowly defined data that compel them to analyze grids, dashboards, and spreadsheets in order to bring about short-term improvements in results.

The brief concludes with twelve recommendations for establishing more effective systems and processes of Data-Driven or Evidence-Informed Improvement and Accountability:

1. *Measure what is valued instead of valuing only what can easily be measured*, so that the educational purposes of schools do not drift or become distorted.
2. *Create a balanced scorecard* of metrics and indicators that captures the full range of what the school or school system values.
3. *Articulate and integrate the components of the DDIA system* both internally and externally, so that improvement and accountability work together and not at cross-purposes.
4. *Insist on high quality data* that are valid and accurate.
5. *Test prudently, not profligately*, like the highest performing countries and systems, rather than testing almost every student, on almost everything, every year.
6. *Establish improvement cultures of high expectations and high support,* where educators receive the support they need to improve student achievement, and where enhancing professional practice is a high priority.
7. *Move from thresholds to growth*, so that indicators focus on improvements that have or have not been achieved in relation to agreed starting points or baselines.
8. *Narrow the gap to raise the bar*, since raising the floor of achievement through concentrating on equity, makes it easier to reach and then lift the bar of achievement over time.
9. *Assign shared decision-making authority, as well as responsibility for implementation, to strong professional learning communities* in which all members share collective responsibility for all students' achievement and bring to bear shared knowledge of their students, as well all the relevant statistical data on their students' performance.

10. *Establish systems of reciprocal vertical accountability,* so there is transparency in determining whether a system has provided sufficient resources and supports to enable educators in districts and schools to deliver what is formally expected of them.

11. *Be the drivers, not the driven*, so that statistical and other kinds of formal evidence complement and inform educators' knowledge and wisdom concerning their students and their own professional practice, rather than undermining or replacing that judgment and knowledge.

12. Create a set of guiding and binding national standards for DDIA that encompass *content standards* for accuracy, reliability, stability and validity of DDIA instruments, especially standardized tests in relation to system learning goals; *process standards* for the leadership and conduct of professional learning communities and data teams and for the management of consequences; and *context standards* regarding entitlements to adequate training, resources and time to participate effectively in DDIA.

Data-Driven Improvement and Accountability

Introduction

Over the past thirty years, two related challenges have spurred educational reform in the U.S. and in many other countries: how to make schools more effective and equitable in their outcomes, and how to make them publicly and politically accountable for delivering those outcomes.[1] Increasingly, there has been a strategic convergence of these two approaches by using more stringent accountability as the prime driver to improve educational performance. There has also been a growing commitment to a particular approach for achieving both improvement and accountability. This has involved collecting, analyzing, and reporting performance data of various kinds at the student, teacher and school levels. Some of the data are used to inform interventions within schools and school systems, and some are used as a basis for evaluating teachers and schools, with the intention of enhancing their effectiveness in promoting student learning.

This policy brief examines the nature, implications and effects of this powerful movement in education: data-driven improvement and accountability (DDIA). It identifies and analyzes the tensions between improvement and accountability in the policies and practices of data use. The brief begins by reviewing the antecedents of DDIA in school improvement efforts, in attempts to quantify school effectiveness, and in system-wide educational reform. It then analyzes the uses of performance data in two other sectors – business and sports – in order to identify generic issues of data-use across sectors, and to highlight what can be gleaned for the benefit of public education from such cross-sector comparisons. Building on this historical and comparative review, the brief then turns to research findings, including our own, on the varying purposes, dynamics and consequences of DDIA in education, particularly, though not exclusively, in the U.S.

The brief concludes by drawing together the implications of the research for DDIA policies and practices that should be advanced in order to maximize their constructive contributions to educational excellence and equity, and it sets out some of the ways in which these findings may be translated into practical and productive legislative provisions.

Effectiveness, Improvement, Accountability & Reform

Beginning in the 1970s and into the 1980s, a substantial and influential body of research in the U.S. and the UK identified the key characteristics of effective schools, defined as those that were successful in producing positive outcomes of achievement, attendance, and behavior.[2] The identified characteristics included having high expectations for students, establishing a safe and orderly school climate, devoting the majority of classroom time to

instruction rather than behavior management, securing parental involvement, setting and checking homework regularly, exercising strong leadership, developing shared decision-making and showing concern for students' welfare. Notably, these characteristics were associated with successful schools serving very different student populations.[3]

Through the 1990s, researchers and system leaders became increasingly aware that, although research could identify the factors that were strongly associated with school effectiveness, there was no clear knowledge base regarding the processes and practices

Reform models were sometimes transferred simplistically from one system to another on much shorter timescales than in the systems in which they had gradually evolved, and they often showed little sensitivity to other differences in culture or political context.

that would enable schools to become more effective over time. A response to this growing concern was the emergence of research on school improvement.[4] Improvement factors highlighted by this research included the importance of developing a shared vision of change,[5] exercising inspiring forms of transformational and instructional leadership,[6] conceptualizing improvement as a continuous process rather than a one-time event, [7] developing cultures of professional collaboration,[8] building professional learning communities,[9] and involving students in the change process.[10] Where schools operated in challenging circumstances such as high poverty and high pupil mobility, high levels of leadership stability and sustainability were also critical for building trust within the schools, as well as with parents and others in the communities.[11]

From the 1990s onwards, findings related to school effectiveness and improvement, such as setting high expectations or standards, establishing a clear focus on instructional improvement, and having strong support for professional development (especially through one-on-one coaching) to help teachers improve their practice, influenced the design and development of large-scale reforms that encompassed entire districts, states and countries.[12]

At the same time, accountability advocates, frustrated by the slow pace and low success rate of existing improvement efforts, advanced a range of new reform strategies that departed from, and in some cases directly contradicted, earlier presumptions about school improvement. For example, the hitherto widely accepted claims that "you cannot mandate what matters to effective practice"[13] and the fact that sustainable improvement requires trust to be built over time, were sometimes sidelined or overturned by other elements of large-scale reform design.[14] Moreover, reform models were sometimes transferred simplistically from one system to another on much shorter timescales than in the systems in which they had gradually evolved, and they often showed little sensitivity to other differences in culture or political context.[15]

Most advocates of large-scale reform in Anglo-American nations have argued that system-wide reforms that result in improved tested achievement, higher graduation rates, and greater rates of college attendance, as well as reductions in the longstanding gaps in these indicators, possess a number of key characteristics. These include a tight focus on two or three key priorities that are clearly measurable (such as raising standards in literacy and math), developing a guiding coalition of key political and other stakeholders who support this priority, setting numerical targets for improvement by specified dates when they should be delivered, exerting strong pressure for change that ranges from relentless implementation in some cases to punitive sanctions for failure to succeed or comply in others, providing and prescribing extensive training, defining and developing detailed curriculum standards and materials, and tracking and monitoring performance in real time at every level from individual students and teachers to the entire system.[16]

Outside the U.S., these strategies of large-scale reform were evident in the education reform measures of the Blair Government in England and Wales,[17] and in the highly publicized reform strategies of Ontario, Canada.[18] In the U.S., through provisions in the No Child Left Behind Act[19] and the Obama Administration's Race to the Top initiatives,[20] the new large-scale accountability and improvement systems have linked indicators based on student test score outcomes to high-stakes consequences, including the reorganization or closure of schools, as well as performance-based pay incentives and sanctions that include dismissal for teachers and principals.

In the main, over more than a decade, these top-down initiatives in the U.S. have not had positive effects on educational excellence or equity.[21] Despite a small number of highly publicized examples, relatively few schools have actually undergone reorganization or closure due to problems of low performance.[22] When measures have been taken to close down ineffective schools, as was the case in Chicago, students were frequently simply transferred to other low performing schools, producing an additional problem; namely, that those schools were outside the students' own neighborhoods.[23] Moreover, because many outcomes-based indicators are crude measures of school effectiveness, they place schools serving high proportions of minority and/or low SES students under such heavy threat of intervention that better qualified teachers tend to gravitate to schools within their district that already post better outcomes, or to other districts that serve more advantaged students and are, therefore, under much less threat.[24]

All this helps to explain why educational inequities in the U.S. show no signs of easing[25] and why a report by McKinsey and Company[26] likens current U.S. educational inequities to the impact of a "permanent national recession". Even large-scale reform models that have been less punitive than those in the U.S., such the educational reform strategy of Ontario, have been more successful in raising the bar of tested achievement for everyone than in narrowing the achievement gap between different social, ethnic or linguistic groups.[27]

One response of policy-makers to the perceived shortcomings of U.S. education has taken the form of outcomes-based,[28] then standards-based, reforms linked to system-wide educational assessments.[29] After mounting evidence that federally imposed performance targets had led some states to redesign their tests, or to modify performance standards, so

that they were easier to pass,[30] an effort led by the Council of Chief State School Officers and the National Governors Association was initiated to devise a set of Common Core State Standards in English/Language Arts and mathematics.[31] These standards have been adopted by 46 states and the District of Columbia. Furthermore, with the active support of the U.S. Department of Education, these two organizations have encouraged states to participate in one of two multi-state consortia that are developing summative assessments aligned to the Common Core State Standards.[32]

The more that the purposes and strategies of educational improvement and accountability have converged, and the more that they have been scaled up from the school, to the district, the state and now the nation, the more central performance data have become. Performance data from all levels are seen as critical to a strategy of

- tracking and monitoring large, complex systems;
- understanding and explaining at a system level what is happening without necessarily having to know the individuals involved;
- helping weaker schools to learn from stronger schools and systems;
- benchmarking schools and systems manifesting sustained success;
- enabling professionals and administrators to monitor the progress of every student and school in real time, in order to make timely interventions so that no child will indeed be left behind.

Burch argues that achievement gaps are technically defined in ways that lead to opportunities for private interventions that have large-scale multinational origins, even if their feel is one of local delivery.[33] She shows that the uses of data, and the concomitant hiring of private organizations to support data analysis, were actually codified under NCLB. Thus, while data-driven improvement and accountability (DDIA) are often presented as contributing to what one of us has called *professional capital* by expecting and enabling teachers to develop and exercise their expertise through participating in collaborative, evidence-informed judgments that support student learning, in practice DDIA is arguably and equally driven by the interests of *business capital* which strives to shape public education reform so as to generate private profit.[34]

Alongside the arguments regarding the intentions and effects of DDIA, there have been efforts to devise technical improvements in the indicators and methodology of DDIA, especially with respect to accountability. The goal has been to obtain more accurate estimates of the relative effectiveness of teachers and schools. A prominent example is the shift from the status-based indicators under NCLB that direct educators towards having their students reach particular thresholds of proficiency, to progress-based indicators of improvement over time, some of which are sensitive to the conditions under which teachers, schools and districts have to operate. These progress-based indicators are often derived from so-called value-added models that take account of differences among classes

and schools in students' prior academic achievement and, sometimes, in demographic characteristics and other contextual factors.[35]

System-wide reform is now the official policy preference for securing both excellence and equity in many or most schools rather than just a few, within timeframes defined as much by election cycles as by those inherent in the work of improvement itself. Strong accountability has assigned urgency and transparency to the effort, and DDIA is now deployed as one of the main tools to engage everyone in the whole change process, all the time. For the most part, system-wide reform in the U.S. has concentrated on schools and school districts, but the reach of the "new accountability" is now extending to teachers and even, in some states, to teacher preparation programs.[36]

For the most part, initiatives undertaken under the aegis of DDIA have maintained a tight focus on school-related data, with an emphasis on threshold targets of cognitive achievement. This kind of focus is indifferent to or dismissive of the substantial research on the association of out-of-school factors with student achievement and underachievement. Such factors include child health, environmental quality, family structure, income and stability, and community resources.[37]

Although there are cases where a few schools and even fewer districts have been able to overcome the immense obstacles that children, families and schools face in very challenging circumstances of poverty and instability, the extensive research findings on the cumulative impact of contextual factors outside the school, and their widespread association with results in educational performance, lead to the reasonable conclusion that achievement gaps will not be substantially reduced unless and until the gaps in other factors are addressed as well.[38] Evidence of the effectiveness of a strategy that addresses the broad range of factors outside, as well as inside, the school that affect student achievement and achievement gaps has been reported by studies in the U.S.[39] and abroad.[40] In the U.S. this has led to the founding of a public advocacy group for adopting a Broader, Bolder Approach[41] to improvement and accountability that adequately encompasses the range of factors that actually have an impact educational progress.

Even if more and better data can be developed as a basis for improvement and intervention in the realms of education, health and social policy, there will still be limits to what DDIA can accomplish. In *Big Data*, Mayer-Schonberger and Cukier (2013) describe many of the ways that data enter into and also enhance our lives. Big data, they say, are about "the ability of society to harness information in novel ways to produce useful insights or goods and services of significant value".[42] They point to how the growing capacity to analyze vast volumes of data has made it possible to predict the spread of flu pandemics, to pinpoint the buildings most likely to be overcrowded fire hazards, and to use complex algorithms to provide just-in-time feedback in response to people's progress in their online learning.

But after their enthusiastic advocacy for Big Data, the authors then advise against turning to data for the answer to all our problems. They stress that there is a "special need to carve out a place for the human: to reserve space for intuition, common sense and serendipity".

"What is greatest about human beings", they say, "is what is precisely what the algorithms and silicon chips don't reveal".[43] This is equally true of those areas of life where problems are pervasive, inequities abound, and human suffering is rampant. Data can help in addressing these issues but in the end, some of our most challenging educational and social problems will not mainly be solved by more or better data, just as they will not be solved by more technology or by any other silver bullet. More and better data can help us make more efficient educational decisions and judgments, but they will not, of themselves, help us make wiser or more humane ones. Often, what we need to alleviate children's suffering and lack of opportunity is not more data or better metrics, but more attention, and more support.

The Uses of Data

The influence and impact of data are not confined to education. In the past two decades, in most of the developed world, data have become an increasingly important and almost inescapable part of all our lives.[44] Data serve many functions, but two are especially important: improvement and accountability.

Data for Improvement

Data can promote improvement in the quality and effectiveness of production and services. For instance, data help companies personalize advertisements and offer products that match customers' digitally tracked preferences and predicted desires. Scanned barcode data enable retailers to employ efficient, just-in-time delivery.[45] For management, digital dashboards indicate when and where performance is strong or weak, guiding interventions to reduce waste and to improve supply chain flow or workforce productivity.

Data usage is also contributing to efforts at improvement in public services.[46] Process and outcomes data identify hospitals that have persistently lower or higher rates of infection, so that poorer performers can learn from their more successful peers.[47] The growing availability of online access to provider ratings also enables users to make more informed choices of service providers and better decisions about their own health.[48]

The use of performance metrics in monitoring government services is seen as essential for both effective management and increased democratic control.[49] A set of indicators that captures key aspects of system functioning gives managers much of the information they need to track, modify or refine the systems for which they are responsible. The general availability of these indicators also contributes to greater transparency for the public that, ideally, provides a basis for developing a general consensus about what is working well and what is not, and that ultimately influences individual decisions on how to vote.[50]

Data for Public Accountability

In addition to fostering improvement, data also increasingly serve an accountability function. This largely takes the form of transparency concerning the standard of services

that are provided and the degree to which they are accomplished by means that are authentic and ethical. Published data on quarterly returns of profits and losses hold companies accountable to shareholders. Rankings on the Dow Jones Sustainability Index or the Reputation Index provide potential investors with data on where companies stand in terms of environmental responsibility and ethical integrity. Published targets push public providers to increase efficiency.

Improvement and Accountability Together

In all sectors, data-driven improvement and accountability (DDIA) is designed to secure improvement in outcomes by promoting accountability through published metrics and increased transparency. The assumption underpinning DDIA is that documented unsatisfactory performance or unethical practices will lead to public outcry and the threat of organizational decline or extinction as people take their business or support elsewhere. Concurrently, the same data provide information that businesses or service providers can use to inform improvement strategies.

But as easily as accountability and the drive for continuous improvement can provoke a constructive sense of urgency, they can also lead to negative consequences. Among other things, data-driven accountability can lead to

- Concentrating on short-term wins at the expense of long-term, sustainable improvement;[51]
- Pushing for higher and higher profit margins at the expense of quality of service to clients;[52]
- Emphasizing standardized actions that lead to easily implemented solutions at the expense of ones that promote greater innovation and creativity;[53]
- Directing efforts to outcomes that are most easily measured or populations that can produce the easiest gains, compared to other outcomes and populations that also have and create value;[54]
- Faking, fabricating and fraudulently representing performance results;[55]
- Investing excessive time and trust in performance metrics and their implications at the expense of energy directed to building relationships within and beyond the organization;[56]
- Creating endemic instabilities by repeatedly opening and closing institutions, and constantly turning over leaders and other staff in order to produce instant turnarounds in measurable results.[57]

To sum up: in other sectors, not just education, the combination of data-driven improvement and accountability can lead to greater quality and integrity, or to deterioration of services and distraction from core purposes. The question addressed by

this brief is what factors and forces can lead DDIA in education to generate more positive and fewer negative outcomes. To do this, we first turn to other sectors in order to examine the two faces of data usage and to illustrate how the tensions characteristic of DDIA manifest themselves in different settings, including the private sphere, rather than being unique to public education.

Data in Other Sectors

There is little empirical work on organizational improvement, effectiveness and accountability across sectors, with most research being theoretical or speculative in nature.[58] However, one of us has co-directed a large-scale study of almost twenty cases of performance beyond expectations in business, sports and education.[59] Performance was measured according to existing metrics available in the different sectors. These included the most obvious metrics of profits, losses and shareholder value in business, league table rankings and cup victories in sports, and examination results or test scores in education. Other measures that were also employed included metrics such as customer satisfaction, reputation, internet "stickiness" and environmental sustainability in business; participation rates in sports; and honors and awards in education. The criterion of performing beyond expectations was judged in relation to previous performance on key indicators, in relation to comparable organizations in the sector, and in relation to contextual factors of unusually daunting challenges or lesser levels of support, such as location in communities of high poverty in education and sports, or low levels of financing and crowd support in sports.

Among the 15 factors that the study found to be associated with high performance, one was the use of performance metrics that were generally regarded as meaningful, fair, broad, varied and sensitively applied among members of the organizations concerned. At the same time, in some of the cases, and in our review of the wider literature on each of the sectors, we also obtained examples of less productive instances of data use. These contrasts offer insights into how data are used -- or could be used -- in education.

DDIA in Sports

The role of performance metrics in improving the results of professional sports teams first achieved broad public awareness with the publication of Michael Lewis's[60] account of how systematic use of player performance statistics raised the performance of the underfunded Oakland Athletics baseball team to World Series standard. Lewis's *Moneyball* describes how the Oakland Athletics became the first baseball franchise to take seriously players' performance statistics as a basis for player recruitment (in contrast to previous reliance on coaches' and scouts' intuitive judgment) and how, as a result, the club was able to make the play-offs year after year, facing teams with triple the payroll.

The use of performance statistics that the Oakland Athletics highlighted, but did not invent, is now standard throughout the sports world. For example, all leading soccer clubs now place video cameras around the pitch, and analyze the movies to derive metrics of

individual and team performance factors (shots made, successful and unsuccessful passes, distances run, goals conceded early or late in the game, and so on), and then have these examined by one or more in-house performance analysts.

In soccer, movement is more flowing and actions are harder to disaggregate than in many stop-start U.S. sports. The difficulties in data-driven improvement in soccer first became evident when Valeri Lobanovsky – manager of the Dynamo Kiev soccer team in the former Soviet Union from the mid-1970s to the early 90s - decided to apply the principles of scientific Marxism to soccer management. Standing by the pitch, he made notations of different moves made by individual players, and connected these to team performance outcomes. When Lobanovsky purchased a large computer, his goal was to combine science and technology to create the perfect soccer team.

Franklin Foer describes how Lobanovsky applied numerical values to every successful and unsuccessful action in the game. The data were put through the computer to produce calculations of "intensivity, activity, error rate", and so on.[61] Lobanovsky was seeking a perfect data-driven system that his players could adopt automatically. He even organized practice matches where players were blindfolded. According to Foer, Lobanovsky's system:

> rewards a very specific style of play: physical and frenetic. Players work tirelessly to compile points. They play defense more aggressively than offense, because that's where points can be racked up. In stifling individual initiative, Lobanovsky's system mimicked the Soviet regime under which it was conceived. Nothing in Lobanovsky's point valuation measures creativity or daring. A vertical pass receives the same grade as a horizontal pass; a spectacular fake means nothing.[62]

The risk of over-reliance on numerical data in soccer still occurs today. The performance analyst at a former English Premier League soccer team pointed to how mechanistic applications of data-driven decision-making in soccer could have perverse effects even in the modern game.[63] He described how in one World Cup of soccer, some managers not only relied on video data to analyze player performance but also placed microchips in their players' boots to gather additional statistics about the number of steps the players took during a game. "There were", he said, "some players who started doing extra steps when the ball went out of play (out of sight of the cameras) so they could up their stats – tell the manager – "Yes, I've done my job this week."

By contrast, this analyst and his club developed a more interactive and collaborative approach to evidence. This included sharing data with the coach; suggesting what it meant in terms of performance or energy levels; inviting players to look at their statistics and see how they compared to average performance levels in the league – and then discussing together specific ways to improve. The analyst described how players who had been skeptical of the system at first, had seen their performance improve after paying attention to their own statistics and the statistics of players they were playing against. After spending many games on the bench, they suddenly started to find that they were being picked for the team every match. "The data *contribute to* rather than *dictating what* they

should do," the analyst commented. "Whether it is technical or tactical, you can have a different interpretation of it." Responding to the data with commands would be sure to fail, because to "prescribe what to do" would "take away the spontaneity and creativity" that accords with the club's philosophy of "freedom of expression".

Performance data at this soccer club were used intelligently, and inclusively; not mechanically and autocratically. In successful sports performance, it seems, data do not automatically drive improvement, but yield evidence that thoughtfully informs it. Data are combined with judgment. They are used to stimulate and support individual and collective responsibility for improvement, not to force or threaten players into compliance. They are connected to the team's creative character and quality, rather than undermining that quality.

DDIA in Business

Nowhere are performance metrics more obviously important than in the profits and losses of business. But the use of performance data in modern business involves far more than quarterly returns or shareholder value alone. Many businesses now employ what is called a triple bottom line, comprising indicators of economic, social and environmental outcomes.[64] This leads to more "balanced scorecards" of organizational performance.[65] At the same time, manufacturers use extensive data at a granular level to identify flaws and possible savings in production and distribution processes that can give companies a profitable edge and also reduce waste in the environment.[66]

In a case study of performance beyond expectations,[67] the incoming CEO of a struggling auto-manufacturer announced that the design-to-production process would have to accelerate from 4 years to an unprecedented period of 18 *months*, and then 15 months, in order to protect the company's survival. Achieving this target without compromising customer safety, a core value of the company, was more important still and would give it an edge over competitors. These targets became the collective responsibility of executive staff, who worked together to brainstorm strategies. Using real-time data, they tagged every indicator as green (on target), red (falling short) and amber (at risk), enabling them to intervene effectively to improve quality and efficiency.

Performance metrics, data dashboards and sophisticated analytics do not always lead to increased improvement in the corporate world, however. Executives and shareholders alike can overreact when one disappointing quarter can be interpreted as signaling the onset of a precipitous decline – leading to what has been called a "bullwhip effect".[68] A quarterly dip may not be an indicator of more serious performance difficulties; rather it could be a statistical anomaly, an unusual season in the sector, a reflection of a wider sector trend, or a short-term effect of leadership turnover or of the temporary disruption caused by implementation of an organizational change.

John Kotter, one of the first advocates of "short-term wins," argued that short-term measurable improvements are valuable because they build confidence, increase credibility among stakeholders, and neutralize cynics. However, Kotter warns, quick wins of a

measurable nature must "make sure that visible results lend sufficient credibility to the (long-term) transformation effort."[69] Failed short-term efforts to secure measurable improvement often do not grasp this fundamental connection.

Cultures of high pressure and high threat to achieve constant short-term gains often lead to negative consequences. Like the ill-fated Everest expedition team of 1996, organizations can too easily become gripped by "summit fever" as they try to reach their targets, with the result that they lose sight of their core purpose and self-destruct in the process.[70] For instance, DeRose and Tichy[71] report how the use of scanning metrics to set through-put targets for supermarket checkout staff led staff at one retailer to fail to help older customers or to make eye contact with shoppers because this would incur time-consuming personal interactions. Enron-like strategies of creative accountancy and outright fraud, or of implementing cost-cutting measures that are environmentally irresponsible or that imperil the safety of workers or customers are the ethical end game of misuses of DDIA.[72]

In the five business cases investigated by Hargreaves & Harris, [73] performance measures and other data metrics were perceived as meaningful, broad and fair; targets were regarded as ambitious but achievable; performance was evaluated over longer as well as shorter time periods; employees were very often involved in setting targets or the targets were at least defined in the context of high-trust relationships; ethics and integrity were high priorities; and all of this supported rather than distracted people from their core mission and purpose. The converse of these findings is that DDIA becomes problematic when measures are narrow in scope and not meaningfully connected to core purposes; when targets are unrealistic and autocratically imposed; and when the organization undergoes what Robert Merton[74] called goal displacement, because numerical targets turn into the new core purpose rather than supporting the business's original and authentic goal.[75]

DDIA in the Public Sector

Although results-based improvement and accountability strategies can achieve positive outcomes, they also raise problems in many sectors, including, as we have seen, business and sports. These issues are also evident in public sector areas such as transportation and health.[76]

Claims about the success of results-based or target-driven improvement in the public sector are widely debated. Nowhere has the debate been more intense than in relation to the success (or not) of target-driven change under the UK's Blair Government – a strategy that Blair's advisor, Sir Michael Barber, terms "Deliverology". A major part of deliverology is the use of data to identify benchmarks either in relation to past performance or in comparison with the performance of peers elsewhere; and to drill down into the actions and interactions of every student, client, teacher, doctor and provider, so that improvement is achieved individual by individual, step by incremental step. In deliverology, data are used to evaluate past and present performance; to pinpoint areas of weakness so they can be addressed in real time;[77] and to set targets and sub-targets that

become "specific measurable commitments"[78] for people at every level of the system, no matter how small.[79] Barber claims that in the UK, data-driven deliverology was successful: "By the end of Blair's second term (2005), around 80% of the ambitious goals we had set out to achieve had actually been achieved. Of the remaining 20% of targets that had been missed, in almost all cases performance had nevertheless improved".[80]

Critics of the target culture, however, pointed out that the imposition of performance targets led to some bizarre consequences. Police officers reduced crime rates by redefining some hard-to-solve crimes as misdemeanours.[81] Hospital staff were reported as meeting their targets for patient waiting times in the emergency room by having ambulances drive the sick and injured around the block until they could be admitted to and processed in the hospital within the targeted time.[82] Before Barber's further development of the target-driven culture of public policy, contracted rail repair workers imperilled passenger safety by mending broken rails within targeted time periods, but neglecting the undergirding ballast and supporting ties that could not be fixed within those periods.[83]

Organizational management expert John Seddon, one of the most vocal critics of policy deliverology, points out that once high-stakes targets with arbitrary numbers are inserted into a system, the system will then organize itself to produce the required result by creating a set of "perverse incentives" to manufacture the desired numerical outcome.[84] These findings are consistent with Campbell's Law (1976),[85] which states that:

> 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.

DDIA in Education

Most of the opportunities and challenges surrounding DDIA in education are similar to those found in other sectors, but with the heightened importance of issues concerning the scope and validity of data and indicators. This section begins with an illustrative example drawn from our recent work in Ontario, Canada, concerning systemic reform strategies in the area of whole school changes that especially impact special education services,[86] and then pinpoints some of the key issues and challenges that emerge in this example that are also evident in the wider research on DDIA.

Ontario is one of four Canadian provinces that scores especially well on the PISA assessments of student.[87] In terms of population, size and heterogeneity of the school population and its mixed economy, Ontario resembles the larger mid-west states. Since 2003, the province has had a reform strategy concentrating on raising the bar of achievement and narrowing achievement gaps in literacy and numeracy. Progress is monitored by means of province-wide standardized assessments (commonly known as the EQAO tests) in Grades 3, 6 and high school.[88] The province established a target of 75% of students meeting or exceeding the threshold cut-score for proficiency (Level 3). The reform strategy has been associated with substantial gains on this indicator, and has had

the additional benefit of raising levels of public confidence in the provincial educational system.[89]

A significant component of Ontario's reform strategy has been data-driven, or evidence-based, improvement and accountability. The EQAO tests are consequential for education professionals as poor results lead to various levels of assistance and/or intervention. However, the threat levels are much lower than in the U.S. in that school performance is not publicly ranked and turnaround strategies rarely involve principal replacement, never lead to school closure, and have no implications for provision of private services such as tutoring support. EQAO test scores are also used in combination with many other indicators of student achievement and school improvement [90] in order to secure external, bureaucratic accountability to system leaders and the wider public on the one hand, and internal, professional accountability (or collective responsibility) among educators on the other.[91]

EQAO data are employed to show year-to year progress of districts and schools towards meeting the 75% threshold target. At the same time, in combination with demographic data, as well as data on school characteristics and populations such as percentages of English language learners, schools can be compared to schools in similar circumstances, in what are termed statistical neighborhoods. These comparisons are made over one-year and three-year time periods (to identify longer term trends so as to discourage bullwhip-like overreactions to short-term shifts) and are also used to identify schools with different profiles of failure or success.[92]

EQAO reports help the system to pinpoint needed support and interventions, and to promote school-to-school learning and inquiry – all within a system-wide drive for organizational improvement that relies, in part, on assessment for learning. In this regard, a key goal is to create a culture of data-use across the system in which strategic decisions at every level are research-based and data-informed.[93] At the district level, part of this data culture involves districts being encouraged to use a wide range of data to set their own targets for improvement.[94]

Our research, which took place from 2009-2011, involved detailed qualitative case studies of ten districts, policy interviews with senior Ministry officials and with designers of the special education strategy, as well as a web-survey of teachers in nine districts that yielded self-reported responses to, and perceptions of, Ontario's reforms that had a bearing on special education issues. The research data pointed to a number of positive features and impacts of DDIA in Ontario's high profile reform effort. Some schools and districts developed effective data cultures in which diagnostic and other data were used to prompt focused conversations about particular children for whom all teachers, across all grades, those with special education responsibilities as well as those in regular classroom roles, held a sense of collective responsibility. In these schools, teachers drew on a wide range of standardized, diagnostic, and other kinds of data such as portfolios and samples of student work to concentrate in a caring and committed way on helping all children as individuals.[95] They "put faces to the data" so that data didn't drive their decisions, but informed their discussions and subsequent interventions in an authentic yet appropriately urgent way.[96]

At the same time, our examination of DDIA in the ten districts, drawn from across the province, revealed several key concerns that have implications for policy, and that are reported elsewhere in the literature on U.S. approaches to DDIA. In a number of cases, especially where school leadership was autocratic or uncertain, or where Ministry intervention staff had been overzealous, there were pressures for teachers to concentrate

> *Cultures of high pressure and high threat to achieve constant short-term gains often lead to negative consequences.*

their attention on "bubble kids";[97] that is, those students whose scores fell in the 2.7-2.9 range, just below the target of Level 3 proficiency This pressure was exerted even though some senior Ministry officials specifically and strongly advised against such practices. In line with the ironic consequences of placing chips in the boots of soccer players to measure their steps, these data indicate that cynical and calculative strategies to raise scores need not be the result of malicious or manipulative policy intent, but can also be the result of the "perverse incentives" described earlier that occur when seemingly arbitrary threshold targets are inserted into a system that exerts pressure from above to reach them within specified time periods.[98]

Survey results for classroom teachers indicated that they were much more likely than administrators or special education resource colleagues to be critical of the EQAO tests. Closed-ended responses indicated much less support for EQAO than for other aspects of the government's reform policy such as its literacy strategy and its support for differentiated instruction. Open-ended responses included teacher reports that EQAO had led to teaching to the test, to special education students being withdrawn from regular classes in order to prepare them for the test, and to a skewed emphasis in the system towards tracking all teachers' performance in order to identify deficiencies and shortfalls in just a few.

When teachers were asked to report where there had been growth in collaborative practices with colleagues, the greatest increases were in those interactions that involved analysis and use of data. At the same time, teachers did not report comparable increases in more traditionally valued collaborative practices such as visiting colleagues' classrooms and joint teaching that have been consistently associated with strong professional cultures as well as gains in student achievement.[99]

One additional issue is that although the province of Ontario has had the twin-pronged reform priority of "raise the bar, narrow the gap", its policies have had more impact on increasing overall test score performance than on narrowing the achievement gap. In particular, while students with special educational needs, and English Language Learners (many of whom are from highly skilled immigrant groups) have shown some gains relative to other students, the effort to narrow the gap has been less successful than overall progress towards reaching thresholds of proficiency. In line with research and analysis by OECD,[100] it is therefore important to consider the alternative strategy of narrowing the gap

in order to raise the bar - i.e. attending to equity in order to increase overall quality, as has been the case in high performing Finland.[101]

Evaluation of DDIA

By widespread agreement,[102] Ontario provides one of the best-case scenarios of how to combine improvement with accountability in ways that include DDIA. It has maintained strong central pressure for both change and transparency of outcomes, but has not instituted strong threats or punitive measures for those who fall short -- on the assumption that poor performance is largely due to insufficient capacity (indicating the need for training and support), rather than to lack of effort or deliberate intransigence.[103] The difficulties that we have uncovered with Ontario's efforts to combine improvement and accountability, alongside those experienced in other sectors, might therefore be expected to be even greater in systems such as those in the U.S. that carry higher threats of sanction and intervention, and that also offer less support for improvement. In most U.S. states, it is outcomes-based accountability rather than school or system improvement that drives DDIA.

In the U.S., in addition to all the contextual factors that threaten student opportunity, achievement and overall wellbeing, such as child poverty, and health and environmental risks,[104] absence of improvement may also signal that the schools or districts concerned do not have the "capacity to build capacity"[105] and to register improvement even when they receive additional resources.[106] Children in the most disadvantaged communities often find themselves not only facing disadvantage and instability at home, but also experiencing an unstable and under-qualified teaching force, high principal turnover, and a politically disruptive environment of constant change, repetitive reforms and school closures that further magnify the insecurities in their lives in school.[107]

In schools and systems that are already short on capacity, more abundant data will be unlikely to help them develop more capacity. Productive use and interpretation of data depend on intelligent leadership, high trust, strong professional relationships and effective collegiality.[108] More and better data can deepen collegiality; but cannot conjure it from nothing.[109] Further, Campbell's Law warns that overreliance on test-based indicators as the predominant form of data use, as is now increasingly common, will likely result in practices and policies that thwart the intentions of DDIA to achieve either real improvement or credible accountability. Both our own research and the associated literature suggest that there are five aspects of an outcomes-based system of DDIA that are essential to evaluations of its impact and effectiveness:

1. The *nature of the data* in type, quality and range;
2. The *indicators* of growth, of progress towards higher standards and threshold targets, and of benchmarked comparisons with peers that are derived from the data;

3. The *interface and interaction* between the data dynamics of accountability and improvement systems respectively;
4. The *consequences* attached to high and low performance; and
5. The *culture and context* of data use.

1. Data

The credibility and utility of a DDIA system depend on the integrity of the data. Accuracy in test scoring, as well as test data recording and transmission, are critical. Accuracy, completeness and recency of administrative data such as school enrollments and student records are equally essential. DDIA in all sectors is also most effective when the data are broad, varied, meaningful, valid, and operate over timescales that provide reliable and stable results. These criteria have been addressed less effectively in public education than in business and sports.

Educators should have varied data for tracking student progress and for pedagogical decision-making. Such data should not only include standardized test scores and off-the-shelf diagnostic assessments, but also teacher-designed assessments, classroom observations, samples of student work in various media, evidence of accomplishments outside school, and so on.[110] This is equivalent to the use of balanced scorecards in business.

Unfortunately, many schools and systems place excessive emphasis on standardized assessments that, because of cost factors and political considerations, do not capture the full spectrum of valued outcomes. At present, tests fail to measure the higher order skills demanded by increasingly rigorous content standards. At a time when the U.S. ranks a lowly 26th out of 29 nations on UNICEF's 2013 indicators of child wellbeing – just three places above bottom country Romania – the neglect of indicators of socio-emotional development is a tragic commentary on the system's priorities. The failure to record and celebrate accomplishments in the arts, creativity, teamwork, facility with digital technologies, and qualities of citizenship also raises questions about the ability of American education systems to attend to the nation's development both as a competitive economy and as a healthy democracy. In line with the predictions of Campbell's Law, outcomes that are not easily specified or measured are given less attention than the drive to improve test scores.

Similarly, if evidence is not collected regarding unintended negative consequences such as a narrowing of the delivered curriculum or outward transfers of students who are thought to imperil the school's ability to meet its performance targets, then school evaluations become fundamentally flawed.[111] Without a balanced scorecard, DDIA in education encourages educators to adopt "perverse" strategies that demonstrate the appearance of satisfactory performance, with adverse consequences for some of the most vulnerable students.

In general, the standardized test scores and other indicators employed in education are not as accurate, relevant or comprehensive as those that are used in business and sports. In this sense, the educational accountability movement has not aligned public education with best business practices but, rather, with a parody of those practices that would not pass muster in any effective business.

To be effectively benchmarked against best practices in business and sports, DDIA should design, select and incorporate a broad range of metrics relating to such outcomes as parental satisfaction, student engagement, and a range of honors and awards. At the same time, just as metrics such as staff retention and leadership stability in business are an important indicator of the likelihood of long term success and sustainability, so too should educational metrics include factors such as working conditions in schools, opportunities for teacher professional development, levels of organizational trust among teachers and with parents and administrators, rates of teacher and principal turnover, levels and appropriateness of teacher certification, and so on. These broader and indeed bolder metrics will also enable system leaders to monitor how their strategies for raising performance on test score data impact other key aspects of the organization's culture such as perceptions of threat, levels of trust, or changes in turnover rates that can affect future performance.

2. Indicators

Separately or in combination, performance results can be used for one or more of three purposes: benchmarking, threshold assessment and measurement of progress or growth.

Benchmarking can be an effective way to improve practice. In industrial benchmarking, businesses do not merely seek to copy others, but strive to learn from them in a deeper way that informs and inspires their own improvement efforts.[112] Benchmarking involves teams scrutinizing their peers, and then learning together what can be adopted and adapted to fit their own organization. In education, variants of industrial benchmarking, such as international benchmarking based on cross-country comparisons of performance, have achieved growing prominence in recent years.[113] According to the Organization for Economic Cooperation and Development,[114] for instance, "disciplined international benchmarking is a common characteristic of the highest-performing countries in education".[115] Many school systems, such as those in England and Ontario, also compare performance across schools, with the goal that poorly performing schools can seek out and secure the assistance of similarly placed but higher performing peers.[116] In states with more equity-oriented funding strategies, comparisons of this kind can also guide reallocation of resources to schools and districts that demonstrate the greatest need.[117]

Unfortunately, comparative benchmarking is sometimes distorted into something more like competitive bench pressing,[118] where school systems and countries treat benchmarking as a kind of Olympic event, focused on overtaking peers, by any means, at almost any cost. In such cases, system administrators and government officials use performance

comparisons to ratchet up a sense of public and political urgency, more than as a starting point to stimulate inquiry and to support genuine improvement.

A second use of test results and other metrics is to demonstrate progress towards a stated threshold or target of excellence or proficiency.[119] On the one hand, aspirational targets can galvanize effort and commitment, especially when they are collectively owned and even collaboratively defined by everyone involved.[120] However, when numerical targets are determined arbitrarily and imposed hierarchically, with punitive consequences mandated for failing to meet them, then calculative practices designed to "game the system" are the predictable and perverse result.

This is evident in the continuing legacy of the No Child Left Behind Act with its Adequate Yearly Progress requirements. In addition to the perverse incentives to give extra attention to those students near the threshold score, other consequences of this reform and its state-level parallels and predecessors have included lowering the proficiency standard, teaching to the test, narrowing the curriculum, and suspending or expelling students who are likely to compromise the desired result.[121]

As the problems that have arisen with indicators linked to threshold measures of performance have become better understood, many systems have turned to using aggregated growth or progress-based measures for school accountability as a supplement to, or a substitute for, threshold measures.[122] Furthermore, the U.S. Department of Education has instituted a waiver program that allows states to submit proposals that would exempt their schools from the NCLB mandate of achieving 100% proficiency by 2014-15.[123] The quid pro quo includes a Federal requirement for introducing progress-based indicators with corresponding targets for improvement. Note that with progress-based indicators, data for each student are aggregated into the calculation of the indicator. Thus, all students contribute directly to the indicator, so there is less incentive to neglect some students in favor of others.[124]

In view of the perverse incentives created by threshold-type indicators, the use of growth or progress-based indicators offers an alternate path for accountability that seems less prone to these distortions. Nonetheless, such indicators are not a silver bullet. The problem with simple growth measures is that observed differences in average growth among schools are likely to be the result not only of differences in true effectiveness, but also of differences in the characteristics of enrolled students (especially in relation to measures of prior academic achievement), the resources available to school staff, and so on. Disentangling the contributions of these confounding factors in order to isolate accurately the differences in school effectiveness is extremely challenging.

To carry out this disentanglement, many states and school systems employ *value-added analysis.*[125] The result of a value-added analysis is the assignment to each school of a measure of its relative effectiveness (in comparison to all other schools) in contributing to its students' learning progress, adjusted for the impact of the confounding factors. This measure is termed the school's value-added estimate.

A typical value-added analysis involves building a statistical model that predicts a student's test score in the current year based on his or her prior test scores and other characteristics (e.g. English language learner, student with disabilities), as well as relevant characteristics of the class and school. The student's contribution to the school's value-added estimate is the difference between the student's actual score and the predicted score. The school's value-added estimate is just the average of these differences. Thus, a school in which students tend to obtain test scores that are greater than predicted will be assigned a positive value-added estimate. Conversely, a school in which the students tend to obtain test scores that are less than predicted will be assigned a negative value-added score.

Although this strategy seems quite reasonable, school value-added estimates can be technically problematic. For example, unlike in a randomized clinical trial, students and teachers do not come together in a class through a proper random mechanism. Due to limitations of the data available, the predictive model cannot completely compensate for the lack of randomness and, therefore, cannot accurately disentangle all the confounding factors in order to isolate the relative effectiveness of different schools.[126]

A second problem is that, because of factors such as small grade-level sample sizes and inaccuracies in scoring systems, schools' value-added estimates can be highly volatile and vary considerably from year to year.[127] In a given year, many "average" schools, and even some considerably above average (based on previous years' results) will suddenly find themselves in the lowest accountability category. The converse is also true. Schools in the lowest category one year can be catapulted into a category above the mean the year after. These seemingly inexplicable but highly consequential fluctuations are not only demoralizing to many school staff, but also damage the credibility of the accountability system as a whole.[128]

Given these difficulties of accuracy and volatility, although schools' value-added estimates contain useful information, they do not constitute a "gold standard" for accountability and, therefore, there is always a need for caution in how they are used and interpreted. All data-based indicators are fallible in one way or another. At the same time, each indicator also draws on different forms of evidence relevant to school effectiveness. For these reasons, it is useful to combine various indicators in order to capture the full range of pertinent evidence. However, no indicator should carry disproportionate weight, as this increases the likelihood of negative consequences accruing in DDIA due to excessive concentration on meeting targets on just one or two indicators.

Within DDIA, strong indicator systems are essential if monitoring and evaluation functions are to operate effectively. Some types and combinations of indicators, we have shown, are more fit for purpose than others. They are more accurate, meaningful and fair and they paint a broader and bolder picture of the practices and performances that stakeholders genuinely value.

Indicator systems can succeed or fail depending on how much technical knowledge and mastery are possessed by the people adopting or using them. The success or failure of

indicator systems also depends on how these systems are used politically. Do they receive proper resource investment? Are they driven by a genuine urge to improve schools or by wanting to keep an eye on the next election? Is their prime intent to build educators' capacities for professional judgment and intervention, or to exert detailed administrative control?

When indicator systems are used for benchmarking, they can stimulate the intelligent professional learning that is essential for school improvement. When they inform and provide instances of aspirational goals, and when they help people appreciate the milestones they can reach along the way, then improvement efforts can be more focused and purposeful. When they assign high priority to progress measures of growth in relation to prior performance, indicator systems can strengthen people's commitment to continuous improvement. However, when political pressures turn intelligent benchmarking into competitive bench-pressing, when threshold achievement measures are arbitrarily defined and hierarchically imposed, and when confidence in value-added growth turns to overconfidence in the measures that underpin it, then improvement becomes the abused or abandoned orphan of an accountability system that has overreached itself and undermined its own purposes.

3. The Interface of Improvement and Accountability

In successful sports teams and businesses, multiple indicators of performance and the conditions that produce it are commonly used to hold individuals, teams and organizations transparently accountable for their performance. These metrics also provide a range of real-time data that enable timely interventions – whether it is altering an offensive strategy on the field or adjusting a production supply chain. In this sense, the twin purposes of improvement and accountability possess a degree of synergy. Nonetheless, in any sector, there are also tensions and even direct conflicts between them.

As we saw earlier, these tensions are most likely to be resolved when there is collaborative involvement in data analysis and collective responsibility for improvement. This kind of resolution has often been difficult to attain in public education settings in the U.S. In part, this has been due to punitive political strategies that have created high threat environments that have undermined collaborative involvement and collective responsibility.

An additional difficulty stems from the fact that the data that are most helpful for improvement are typically not the data most commonly employed for accountability purposes. Teachers and principals working on improvement require timely information about individual students, classes and departments so they can devise and implement appropriate interventions. Throughout the school year, teachers collect and process quantitative and qualitative data that are more useful to them than the information provided by test results from end-of-year summative assessments – results that often arrive in the summer when it is too late to apply what might be learned to the students who were tested. Moreover, the data that are most useful for teachers in helping individual

students are diagnostic data concerning how these students performed on particular questions or items. These kinds of data can pinpoint where students are experiencing difficulties and enable to teachers to make precise, just-in-time interventions.[129] In high stakes, standardized testing, however, item-level performance data is often unavailable due to item release policies that are designed to protect confidentiality and, therefore, also the credibility of the scores. The consequence is that teachers and students expend enormous quantities of time preparing for tests that produce results that have little or no value for the students who took them or for the teachers of those students.

In contrast to diagnostic data, data for accountability purposes operate at the level of a school, district or a state. In the interest of fairness, accountability systems demand "comparable results", usually obtained from standardized assessments, that are centrally

> *The high-stakes, high-pressure environment of educational accountability in the U.S, in which arbitrary numerical targets are hierarchically imposed, has led to extensive gaming and continuing disruptions of the system with unacceptable consequences for the learning and achievement of the most disadvantaged students.*

prepared and scored, and also confidential (to prevent cheating). The validity of these summative assessments rests on the extent to which they fully represent the scope of the content standards established by the state and on whether specific subpopulations of students such as English Language Learners or students with disabilities are not disadvantaged by the nature and format of the assessment.[130]

In practice, it is extremely difficult to create valid standardized assessments that are fully aligned with rich content standards and challenging performance standards. Since test designers operate under severe, state-mandated constraints of time (the tests typically occupy only one or two class periods) and cost (especially in connection with human scoring), they must make tradeoffs that compromise full validity. These tradeoffs can include giving less attention to certain standards, particularly those requiring extended responses. They can also involve reusing items from previous administrations, which can lead to score inflation as teachers become more familiar with the item formats or even the content of specific items.[131]

To sum up, there are at least two tensions between improvement and accountability in DDIA. First, there is a tension between local data used for diagnosis and remediation and comparable, system-level data used for accountability purposes. Second, there is a tension between properly assessing high level learning goals and the constraints of cost and time that govern the design and administration of standardized assessments. Several strategies can be adopted in response to these tensions:

- Increase the expenditures allocated to the development and processing of standardized assessments in order to enhance their validity;

- Improve the cost-effectiveness of standardized assessments by reducing the frequency of testing and by shifting from testing a census of all students, to testing representative samples of students.

- Assign some weight in accountability to indicators based on teacher-generated data and teacher-designed assessments (assuming that they are subject to periodic external audit) such as those now being developed in the high performing Canadian province of Alberta as an alternative to the provincial standardized achievement tests that the province has decided to abolish;

- Discourage school rankings based on simple ladders of achievement and improvement by creating richer school profiles or balanced scorecards that include both standardized test scores and a range of school performance indicators.

In summary, in technical design terms, DDIA in education works best if it incorporates a broad and coherent system of formative, interim and summative assessments, and if it acknowledges and addresses the inevitable tensions between improvement and accountability.

4. Consequences

The impact of DDIA is profoundly influenced by the consequences that flow from it, as well as by how these consequences affect different classes of participants – teachers, students and administrators - who are engaged in or are affected by DDIA. Three considerations are critical: differentiation, timescale, and magnitude.

With regard to the problem of differentiation, Elmore[132] notes that when the stakes attached to test scores differ for educators and for students, distortions in the results may follow. For example, if the stakes for students are relatively low, many students will exert less than maximal effort and, as a result, will underperform. On the other hand, if the stakes for teachers and schools are relatively high, they will certainly feel the pressure to secure the best results possible. This mismatch can lead some teachers to provide inappropriate support to their students during the test administration or even to fabricate test results.

A second issue concerns the timescale over which the consequences are applied. When consequences mainly depend on the most recent evaluations (and if those evaluations are, in turn, heavily reliant on current data), and on the volatile swings in results from one year to the next that often occur, this leads to lurches between sanctions and rewards that undermine the credibility of the whole system. Similarly, if negative consequences such as school closures or the firing of teachers or leaders are applied with such haste that there is neither respect for due process nor opportunity to return to good standing, then perceptions of lack of fairness will breed cynicism and resistance and derail the quest for meaningful school improvement.

The last issue concerns the magnitude or severity of the consequences. A principal tenet among U.S. policy makers today is that for an educational accountability system to have the desired impact, it must result in significant consequences. This belief is at odds with much of the research in education and in other sectors,[133] which shows that large, extrinsic rewards can dampen intrinsic motivation and that tryouts of such reward systems yield minimal to no improvement. The belief in the necessity of significant consequences is also out of step with the accountability practices of high performing countries such as Canada, Finland and Singapore that do not attach external rewards or punitive consequences to the extremes of performance on achievement tests.[134]

Despite the clear evidence from international benchmarking and educational research, U. S. states still implement accountability systems with high stakes consequences, with predictable results. For example, Daly and his colleagues (2011)[135] conducted research in 549 California schools on how educators perceived and responded to high threats of sanctions under No Child Left Behind. They found that principals in schools officially under notice to improve were more likely than their counterparts to experience difficult communication with their districts, to have developed lowered self-efficacy in believing they could lead improvement and, as a result, to adopt sub-optimal strategies like concentrating on the students near the cut scores. Setting "stretch" goals and threatening punitive consequences without providing adequate support, Daly and his colleagues conclude, reinforces transactional leadership practices that focus on reaching narrowly defined, short-term targets.

In short, policy makers' preferred strategy of imposing high stakes consequences on educators in order to "get their attention" is at odds with a great deal of empirical research. Moreover, educators' perceptions of fairness also depend on their views of the completeness of the underlying data and the fairness of the constellation of indicators. Other than in the most severe cases where basic safety is at issue, and in line with the practices of high performing systems, policy makers should avoid responding to "negative" data with premature interventions and the associated risks of succumbing to the bullwhip effect. Instead, they should commence by inquiring into the accuracy and meaning of the data, follow up with swift and sure support where underperformance has been affirmed, and make closure or top-down intervention a remedy of last resort.

5. Culture and Context

No matter how plentiful the metrics available for data-driven improvement may be, they will have little effect unless educators have not only the human capital and resources to analyze and act upon them effectively, but also the social capital to collaborate in high-trust teams with collective commitments to continuous improvement.[136] A key study by Datnow and colleagues (2007)[137] of four districts judged to be making effective use of data for improvement highlights six factors that are important in a culture of DDIA:

- Defining a small number of clearly focused goals that concentrate teachers' data-driven improvement efforts;

- Creating a culture in which data are valued in helping solve improvement questions and where there is reciprocal accountability between schools and the central office, in an environment characterized by trust rather than threat;
- Investing in a functional data management system and a staff of specialists responsible for access and development, as well as for expert data analysis when needed;
- Compiling and archiving data that are linked to standards and priorities, yet also varied and balanced in nature;
- Providing time (through classroom coverage) for professional development and assistance from outside experts and from other schools in order to develop teachers' professional capital in all facets of assessment literacy;
- Supplying tools that enable data-informed, just-in-time feedback to guide teachers' pedagogical decisions.

Research that examines both more and less successful instances of DDIA, shows that well-led, high-trust environments focused on using data for continuous improvement, rather than to escape threatened sanctions, are critical to achieving sustainable success.[138] When systems set high but attainable expectations and make adequate resources available, when they do this as part of a process of shared goal-setting, and when they balance formal data with experiential judgment, then educators can be motivated to assume collective responsibility for the school's success, and to adopt new strategies as well as allocate resources based on considered decisions informed by a range of evidence. However, an excessive focus on data-driven collaboration can displace and downgrade other, equally valuable, forms of professional collaboration, such as team teaching and curriculum planning. Professional cultures should appreciate and honor multiple forms of collaboration in the drive for improvement.[139] With respect to how and how much they use data in relation to professional decision-making, educators should be the drivers, not the driven.

Recommendations

Contrary to the practices of high performing countries on international assessments, and of high performing organizations in business and sports, DDIA in the U.S has been skewed towards accountability over improvement. Indicators, metrics and targets have been narrow rather than broad, defined inaccurately and applied in problematic ways. Test score data have been collected and reported over too short timescales that make them unreliable for accountability purposes, or reported long after the student populations to which they apply have moved on so that they have little or no value for improvement purposes. DDIA in the U.S. has focused on what is easily measured rather than on what is educationally valued. It holds schools and districts accountability for effective delivery of results without holding system leaders accountable for providing the resources and conditions that are necessary to secure those results.

The high-stakes, high-pressure environment of educational accountability in the U.S, in which arbitrary numerical targets are hierarchically imposed, has led to extensive gaming and continuing disruptions of the system with unacceptable consequences for the learning and achievement of the most disadvantaged students. These perverse consequences include loss of learning time by repeatedly teaching to the test; narrowing of the curriculum to that which is easily tested; devoting undue attention on "bubble" students near the threshold target of required achievement at the expense of high-needs students whose current performance falls further below the threshold; constant rotation of principals and teachers in and out of schools where students' lives already have high instability in order to meet the pressure for short-term results; and, in the most egregious instances, criminally culpable cases cheating.

Last, when accountability is prioritized over improvement, DDIA neither helps educators make better pedagogical judgments nor enhances educators' knowledge of and relationships with their students. Instead of being informed by the evidence, educators become driven to distraction by narrowly defined data that compel them to analyze dashboards, grids and spreadsheets in order to bring about short-term improvements in results.

Our research findings, as well as those in the literature, point to predictable successes and shortcomings of DDIA, depending on how it is designed and implemented. These have significant implications for education policy. Together, the following 12 recommendations that are derived from our analysis of the relevant research can provide a foundation for a coherent strategy that could help the U.S. turn DDIA on to a more productive course.

1. ***Measure what is valued instead of valuing only what can easily be measured.*** This tenet should be as strong in education as it is in business and sports. Metrics and indicators should accurately reflect the range and levels of the learning goals and other priorities set by the state such as critical reasoning, emotional and social learning, creativity and teamwork.

2. ***Create a balanced scorecard.*** Collect evidence on a regular schedule from different sources to capture different aspects of system functioning and multiple student outcomes. The student data and administrative data that are routinely collected and reported in most school systems do not render a sufficiently complete picture of the education that students receive, nor of the factors that affect students' learning. Balanced scorecards should include, but not be restricted to, the time allocated to each subject by grade, suspension rates, staff turnover rates, teacher absenteeism, diagnostic assessments, survey results of student engagement, teacher certification, student mobility, and so on.

3. ***Articulate and integrate the components of the DDIA system both internally and externally.*** Internally, different data types (e.g. formative, interim and summative assessments) and their use should complement rather than contradict one another. For this reason, all assessments should be coherent with a common set of content and performance standards. Externally, DDIA should cohere with other parts of the improvement and accountability system. For example, efforts to strengthen professional collaboration will be stymied by reward systems that are driven by indicators of individual teacher effectiveness.

4. ***Insist on high quality data.*** Institute a regular and rigorous quality assurance audit of all indicators used for improvement and accountability. In particular, test-based indicators used for high stakes decisions should meet industry standards with respect to accuracy, reliability, year-to-year stability, and validity.

5. ***Test prudently, not profligately.*** One of the objections to increasing the level of sophistication of tests and indicators is the increased cost. But it is counterproductive to control costs by settling for lower test quality that impedes improvement, diminishes authentic accountability, and undermines the system's credibility. A widely used and successful alternative is to reduce the scope and frequency of testing. This can be achieved by testing at just a few grade levels (as in England, Canada and Singapore), rather than at almost every grade level. Another option is to test a statistically representative sample of students for monitoring purposes (as in Finland), rather than a census of all students. Yet another route is to test different subjects in a rotating cycle (e.g. math is centrally tested and scored once every 3 or 4 years), with moderated teacher scoring of assessments occurring during the intervening years (as in Israel). All these options lower the costs of testing and create opportunities for compensatory improvements in quality. At the same time, not testing all students, every year, reduces the perverse incentives to teach to the test and to concentrate disproportionately on easily "passable" students.

6. ***Establish improvement cultures of high expectations and high support.*** Set challenging performance standards for students and attainable benchmarks for schools, with the proviso that adequate support for continuous school improvement will be provided by the system.[140]

7. ***Move from thresholds to growth.*** Systems should limit the use of imposed numerical targets tied to threshold criteria as these induce a host of perverse incentives. Indicators based on student progress, by comparison, encourage educators to address the needs of all students and to keep moving forward without the anxiety about reaching one particular target at a specified time. When targets are retained, they should be seen as fair, which will be more likely if they are established within a high-trust system and set collaboratively by professionals.[141]

8. ***Narrow the gap to raise the bar.*** Test score gaps reflect, in large part, differences in family and community assistance available to students, as well as differences in the levels of resources and capacity within schools and school districts. Evidence of such disparities should trigger support for both students and schools. International evidence indicates that in education, quality cannot be achieved without equity.[142] The pre-requisite for raising the bar is narrowing the gap. Bringing up the floor brings people closer to lifting the ceiling.

9. ***Assign shared decision-making authority, as well as responsibility for implementation, to strong professional learning communities.*** The DDIA system should support high trust professional communities characterized by collective responsibility for all students' success, and in which data-informed discussions are valued alongside other effective modes of professional collaboration. Mutual support among educators then becomes the norm and students are less likely to "fall through the cracks" as they move from one class to another. High-trust environments assign significant authority to professional communities for shared decision-making in relation to data-informed judgments.

This limits the exercise of capricious authority by administrators based on privileged interpretations of data.

10. ***Establish systems of reciprocal vertical accountability.*** Complementing lateral processes of collective responsibility, systems of reciprocal or mutual vertical accountability encourage and require individuals at all levels to carry through needed actions, establish proper conditions and supports, maintain productive professional relationships, and behave with integrity and respect. Reciprocal accountability can be monitored through 360 degree evaluations, audits of relational trust levels among all the system's members, and peer-based involvement in system decisions about competence and performance.

11. ***Be the drivers, not the driven.*** Data are neither a substitute nor a surrogate for professional judgment. The purpose of data is to support, stimulate and inform the judgment that is necessary for educational improvement and accountability. Expertise has no algorithm. Wisdom does not manifest itself on a spreadsheet. Numbers must be the servant of professional knowledge, not its master. Educators can and should be guided and informed by data systems; but never driven by them.

12. ***Create a set of guiding and binding national standards for DDIA.*** These should comprise content standards for accuracy, reliability, stability and validity of DDIA instruments, especially standardized tests in relation to system learning goals; process standards for the leadership and conduct of professional learning communities and data teams and for the management of consequences; and context standards regarding entitlements to adequate training, resources and time to participate effectively in DDIA.

Conclusion

Debates regarding the future of data-driven improvement and accountability in the U.S. will not be about whether public education should or should not be data-driven or evidence-informed. If rich information can be made available to help all stakeholders make better judgments about, and provide improved support for, all students, then professionally and ethically, that information should not be ignored. The more important question is how to capitalize on the positive potential of DDIA without falling victim to its weaknesses. The essential choice is whether DDIA in public education will become as autocratic and mechanistic as Lobanovsky's Cold War soccer system, or whether it will be used to enhance and enrich the quality of collective professional judgment so that all America's students will reap the benefits of a better education.

Notes and References

1 Elmore, R.F. (2004). *School reform from the inside out: Policy, practice, and performance*. Cambridge, MA: Harvard Education Press;Daly, A. (2009) Rigid response in an age of accountability: The potential of leadership and trust. *Educational Administration Quarterly, 45*(2), 168-216.

2 Student achievement was typically measured by standardized instruments of cognitive attainment.

Edmonds, R. (1979). Effective schools for the urban poor. *Educational Leadership, 37*(1), 15-18;

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Shepton Mallet, UK: Open Books;

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988) *School matters: The junior years*. Shepton Mallett, UK: Open Books;

Levine, D.U., & Lezotte, L.W. (1990). *Unusually effective schools: A review and analysis of research and practice*. Madison, WI: National Center for Effective Schools Research and Development;

Sammons, P. (1999). *School effectiveness: Coming of age in the twenty-first century*. Lisse: Swets & Zeitlinger Publishers.

3 Smith, D. J. & Tomlinson, S. (1989). *The school effect: A study of multi-racial comprehensives*. London: Policy Studies Institute.

4 Reynolds, D., Creemers, B., Bollen, R., Hopkins, D., Stoll, L., & Lagerwijs, N. (1996). *Making good schools: Linking school effectiveness and improvement*. London: Routledge.

5 Stoll, L. & Fink, D. (1996). *Changing our schools: Linking school effectiveness and school improvement*. Buckingham, England: Open University Press.

6 Leithwood, K., Jantzi, D., & Steinbach, R. (1999). *Changing leadership for changing times.* Florence, KY: Taylor and Francis Group;

Hallinger, P. & Heck, R.H. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1995. *Educational Administration Quarterly*, *32*(1), 5-44.

7 Louis, K.S., & Miles, M.B. (1990). *Improving the urban high school: What works and why*. New York: Teachers College Press.

8 Fullan, M. & Hargreaves, A. (1996). *What's worth fighting for in your school?* (Revised ed.). New York: Teachers College Press.

9 Hord, S.M. (1997). *Professional learning communities: Communities of continuous inquiry and improvement*. Austin, TX: Southwest Educational Development Laboratory;

Newmann, F. & Wehlage, G. (1995). *Successful school restructuring*. Madison, WI: Center on Organization and

Restructuring of Schools;

McLaughlin, M.W. & Talbert, J.E. (2001). *Professional communities and the work of high school teaching*. Chicago IL: University of Chicago Press.

10 Rudduck, J. & Flutter, J. (2000). Pupil participation and pupil perspective: 'carving a new order of experience'. *Cambridge Journal of Education*, *30*(1), 75-89.

11 James, C. (2006). *How very effective primary schools work*. London: Paul Chapman Publishing;

Hargreaves, A. & Fink, D. (2006). *Sustainable leadership*. San Francisco: Jossey-Bass;

Mintrop, H. (2004). *Schools on probation: How accountability works (and doesn't work).* New York: Teachers College Press;

Bryk, A.S. & Schneider, B.L. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage Foundation Publications;

Harris, A. & Chapman, C. (2002). Leadership in schools facing challenging circumstances. *Management in Education*, *16*(1), 10-13;

Thomson, P. (2002). *Schooling the rustbelt kids: Making the difference in changing times*. Sydney, Australia: Allen & Unwin Academic;

Murphy, J. & Meyers, C.V. (2007). *Turning around failing schools: Leadership lessons from the organizational sciences*. Thousand Oaks, CA: Corwin.

12 Slavin, R.E. (1996). *Every child, every school: Success for all*. Thousand Oaks, CA: Corwin;

Fullan, M. (2000). The return of large-scale reform. *Journal of Educational Change*, *1*(1), 5-27;

Crévola, C.A. & Hill, P.W. (1998). *Class: Children's literacy success strategy, an overview*. East Melbourne, Australia: Catholic Education Office;

Hill, P.W. & Crévola, C.A. (1999). Key features of a whole-school, design approach to literacy teaching in schools. *Australian Journal of Learning Difficulties*, *4*(3), 5-11;
Elmore, R.F. & Burney, D. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District# 2, New York City*. New York: National Commission on Teaching & America's Future.

13 Berman, P. & McLaughlin, M.W. (1976). Implementation of educational innovation. *The Educational Forum, 40*(3), 345-370.

14 Daly, A. (2009) Rigid response in an age of accountability: The potential of leadership and trust, *Educational Administration Quarterly, 45*(2), 168-216.

15 Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey-Bass;

Hargreaves, A. (2009). Change from without: Lessons from other countries, systems, and sectors. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second international handbook of educational*

change (105-117). Dordrecht, Netherlands: Springer;

Stein, M.K., Hubbard, L., & Mehan, H. (2004). Reform ideas that travel far afield: The two cultures of reform in New York City's District# 2 and San Diego. *Journal of Educational Change*, *5*(2), 161-197.

16 Elmore, R.F., & Burney, D. (1997). *Investing in teacher learning: Staff development and instructional improvement in Community School District# 2, New York City*. New York: National Commission on Teaching & America's Future;

Barber, M. (2007). *Instruction to deliver: Tony Blair, public services and the challenge of achieving targets*. London: Politico's Publishing;

Fullan, M. (2004). *Leadership & sustainability: System thinkers in action*. Thousand Oaks, CA: Corwin;

Hopkins, D. (2007). *Every school a great school: Realizing the potential of system leadership*. Buckingham, UK: Open University Press.

17 Barber, M. (2009). From system effectiveness to system improvement. In A. Hargreaves & M. Fullan (Eds.), *Change Wars* (71-94). Bloomington, IN: Solution Tree.;

Hopkins, D. (2009). Every school a great school: Realising the potential of system leadership. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.) *Second International Handbook of Educational Change* (741-764). Dordrecht, Netherlands: Springer.

18 The Organisation for Economic Co-operation and Development (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: Author;

Fullan, M. (2005). *Leadership & sustainability: System thinkers in action*. Thousand Oaks, CA: Corwin;

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company. Retrieved May 30, 2013, from http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools/.

19 No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

20 U.S. Department of Education (2009, November). *Race to the Top program executive summary*. Washington, DC: Author. Retrieved October 19, 2013, from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

21 Sunderman, G.L. (Ed.) (2008). *Holding NCLB accountable: Achieving accountability, equity, & school reform*. Thousand Oaks, CA: Corwin;

Braun, H.I., Chapman, L. & Vezzu, S. (2010) The black-white achievement gap revisited. *Education Policy Analysis Archives*, 18(21).

22 Steiner, L. (2009). *Tough decisions: Closing persistently low-performing schools*. Lincoln, IL: Center on Innovation and Improvement. Retrieved May 26, 2013, from http://www.centerii.org/survey/downloads/Tough_Decisions.pdf.

23 De la Torre, M. & Gwynne, J. (2009). *When schools close*. Chicago, IL: Consortium on Chicago School Research.

24 Ladd, H.F. (2008). Teacher effects: What do we know? In G. Duncan & J. Spillane (Eds.), *Teacher quality: Broadening and deepening the debate*. Evanston, IL: Northwestern University. Retrieved June 2, 2013, from http://tqn.sesp.northwestern.edu/.

25 Equity and Excellence Commission. (2013). *For each and every child*. Washington, DC: Author.

26 Auguste, B.G., Hancock, B., & Laboissière, M. (2009). The economic cost of the US education gap. *The McKinsey Quarterly*, June.

27 Hargreaves, A. & Braun, H. (2012). *Leading for all: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

28 Spady, W.G. (1994). *Outcome-based education: Critical issues and answers*. Arlington, VA: American Association of School Administrators.

29 Smith, M.S., & O'Day, J. (1990). Systemic school reform. *Journal of Education Policy*, *5*(5), 233-267.

30 United States Government Accountability Office (2009). *No Child Left Behind Act: Enhancements in the Department of Education's review process could improve state academic assessments*. Washington DC: U.S. Government Printing Office.

31 National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Author.

32 The two consortia are the Partnership for the Assessment of Readiness for College and Career (PARCC) and the Smarter Balanced Assessment Consortium (SBAC).

33 Burch, P. (2009). *Hidden markets*. New York: Routledge.

See also:
Ball, S.J. (2007). *Education plc: Understanding private sector participation in public sector education*. New York: Routledge.

34 Hargreaves, A. & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. New York: Teachers College Press.

35 Hargreaves, A. & Harris, A. (2011). Performance beyond expectations. Nottingham, UK: National College for Schol Leadership;

Hout, M. & Elliott, S.W. (Eds.). (2011). *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.

36 Center for American Progress (2011). Getting better at teacher preparation and state accountability. Retrieved May 26, 2013, from http://www.americanprogress.org/issues/2012/01/pdf/teacher_preparation.pdf;

Gansle, K.A., Noell, G.H., Knox, R.M., & Schafer, M.J. (2010). Value-added assessment of teacher preparation in Louisiana: 2005-2006 to 2008-2009. Baton Rouge, LA: Louisiana State University;

Hanushek, E.A., & Rivkin, S.G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics, 4*(1), 131-157;

Cochran-Smith, M., Cannady, M., McEachern, K., Mitchell, K., Piazza, P., Power, C., & Ryan, A. (2012). Teachers' education and outcomes: Mapping the research terrain. *Teachers College Record*, *114*(10).

37 Duncan, G.J. & Murnane, R.J. (2011). *Whither opportunity?: Rising inequality, schools, and children's life chances*. New York: Russell Sage Foundation Publications;

Berliner, D. (2006). Our impoverished view of educational reform. *The Teachers College Record*, *108*(6), 949-995.

38 Barton, P. & Coley, R. (2009). *Parsing the achievement gap II. Policy Information Report*. Princeton, NJ: Policy Information Center, Educational Testing Service;

Rothstein, R. (2004). *Class and schools: Using social, economic and educational reform to close the Black-White achievement gap*. Washington DC: Economic Policy Institute.

39 Henig, J.R. (2012). The politics of data use. *Teachers College Record*, *114*(11).

40 Sahlberg, P. (2011). *Finnish lessons*. New York: Teachers College Press.

41 Broader, Bolder Approach (2013) Retrieved April 23, 2013, from http://www.boldapproach.org/.

42 Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt, 2.

43 Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt, 196.

44 Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Boston: Houghton Mifflin Harcourt.

45 Harvey, D. (1989). *The condition of postmodernity* (Vol. 14). Oxford, UK: Blackwell.

46 Leadbeater, C. (2004). *Personalisation through participation: A new script for public services*. London: Demos.

47 Pronovost, P.J., Schumacher, K., Berenholtz, S.M., Lubomski, L.H., Watson, S., Colantuoni, E., et al. (2010). Sustaining reductions in catheter related bloodstream infections in Michigan intensive care units: Observational study. *British Medical Journal*, *340*;

Gawande, A. (2007). *Better: A surgeon's notes on performance*. New York: Picador.

48 Leadbeater, C. (2004). *Personalisation through participation: A new script for public services*. London: Demos.

49 Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T. & Smith, P. (2005). Performance indicators: Good, bad and ugly. *Journal of the Royal Statistical Society*: *Series A,* 168, Part 1.

50 Yankelovich, D. (1991). *Coming to public judgment: Making democracy work in a complex world.* Syracuse, NY: Syracuse University Press.

51 Kotter, J.P., & Cohen, D.S. (2002). *The heart of change: Real-life stories of how people change their organizations*. Boston, MA: Harvard Business Press.

52 Bevan, J. (2001). *The rise and fall of Marks & Spencer*. London: Profile Books.

53 One critique of the literature on the growth of knowledge societies is that the claims regarding the need for 21st century skills are exaggerated and that many service sector jobs such as banking work are becoming increasingly standardized and affording reduced opportunities for discretionary judgment. See Crawford, M. B. (2009). *Shop class as soulcraft: An inquiry into the value of work*. New York: Penguin Press.

54 Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231-268.

55 Brewster, M. (2003). *Unaccountable: How the accounting profession forfeited a public trust*. Hoboken, NJ: Wiley.

56 Seldon, A. (2009). *Trust: How we lost it and how to get it back*. London, UK: Biteback Publishing.

57 Mintrop, H. (2004). *Schools on probation*. New York: Teachers College Press.

58 Collins, J. (2001). *Good to great: Why some companies make the leap... and others don't*. New York: Harper Business.

59 Hargreaves, A. & Harris, A. (2011). *Performance beyond expectations*. Nottingham, UK: National College for School Leadership.

60 Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York: W.W. Norton.

61 Foer, F. (2004). *How soccer explains the world: An unlikely theory of globalization*. New York: HarperCollins, 159.

62 Foer, F. (2004). *How soccer explains the world: An unlikely theory of globalization*. New York: HarperCollins, 160.

63 Hargreaves, A. & Harris, A. (2011). *Performance beyond expectations*. Nottingham, UK: National College for School Leadership;

Aspects of this case were first reported in Hargreaves, A., & Shirley, D. (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, CA: Corwin.

64 Elkington, J. (1997). *Cannibals with Forks: The Triple Bottom Line of 21st Century Business*. London: Capstone.

65 Kaplan, R.S. & Norton, D.P. (1996). Using the balanced scorecard as a strategic management system. *Harvard Business Review*, *74*(1), 75–85.

66 McDonough, W. & Braungart, M. (2002). *Cradle to cradle: Remaking the way we make things*. New York: North Point Press.

67 Hargreaves, A. & Harris, A. (2011). *Performance beyond expectations*. Nottingham, UK: National College for School Leadership.

68 Sheffi, Y. (2005). *The resilient enterprise: Overcoming vulnerability for competitive advantage*. Cambridge, MA: MIT Press.

69 Kotter, J.P. (1996). *Leading change*. Boston, MA: Harvard Business Press, 125.

70 Kayes, D.C. (2006). *Destructive goal pursuit: The Mount Everest disaster*. Basingstoke, UK: Palgrave Macmillan.

71 DeRose, C. & Tichy, N.M. (2012). *Judgment on the front line: How smart companies win by trusting their people*. New York: Portfolio/Penguin.

72 Brewster, M. (2003). *Unaccountable: How the accounting profession forfeited a public trust*. Hoboken, NJ: Wiley;

Ordóñez, L.D., Schweitzer, M.E., Galinsky, A.D., & Bazerman, M.H. (2009). Goals gone wild: The systematic side effects of overprescribing goal setting. *The Academy of Management Perspectives*, *23*(1), 6-16;

Wolmar, C. (2001). *Broken rails: How privatisation wrecked Britain's railways*. London, UK: Aurum.

73 Hargreaves, A. & Harris, A. (2011). *Performance beyond expectations*. Nottingham, UK: National College for School Leadership;

Hargreaves, A. & Shirley, D. (2009). *The fourth way: the inspiring future for educational change*. Thousand Oaks, CA: Corwin.

74 Merton, R. (1957). *Social theory and social structure* (revised ed.). Glencoe, IL: The Free Press.

75 Burkeman, O. (2012). *The antidote: Happiness for people who can't stand positive thinking*. New York: Faber and Faber.

76 Lebow, R. & Spitzer, R. (2002). *Accountability: freedom and responsibility without control*. San Francisco, CA: Berrett-Koehler;

Rothstein, R. (2004). *Class and schools: Using social, economic and educational reform to close the black-white achievement gap*. Washington DC: Economic Policy Institute.

77 Barber, M. & Moffit, A. (2011). *Deliverology 101: A field guide for educational leaders*. Thousand Oaks, CA: Corwin Press.

78 Barber, M. (2007). *Instruction to deliver: Tony Blair, public services and the challenge of achieving targets*. London: Politico's Publishing, 50.

79 Barber, M. (2007). *Instruction to deliver: Tony Blair, public services and the challenge of achieving targets*. London: Politico's Publishing, xix.

80 Barber, M. & Moffit, A. (2011). *Deliverology 101: A field guide for educational leaders*. Thousand Oaks, Calif.: Corwin Press, viii.

81 BBC News (2007). Targets 'destroy trust in police.' *BBC News - Home*. Retrieved May 31, 2013, from http://news.bbc.co.uk/2/hi/uk_news/7145860.stm.

82 Seddon, J. (2008). *Systems thinking in the public sector: The failure of the reform regime… and a manifesto for a better way*. Axminster, UK: Triarchy Press.

83 Wolmar, C. (2001). *Broken rails: How privatisation wrecked Britain's railways*. London, UK: Aurum.

84 Seddon, J. (2008). *Systems thinking in the public sector: The failure of the reform regime… and a manifesto for a better way*. Axminster, UK: Triarchy Press.

Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T., & Smith, P. (2005). Performance indicators: good, bad and ugly. *Journal of the Royal Statistical Society: Series A*, 168, Part 1.

85 Campbell, D.T. (1976). *Assessing the impact of planned social change*. Kalamazoo, MI: Evaluation Center, College of Education, Western Michigan University.

86 Hargreaves, A. & Braun, H. (2012). *Leading for All: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

87 Knighton, T., Brochu, P., & Gluszynski, T. (2010). *Measuring up: Canadian results of the OECD PISA study*. Ottawa: Statistics Canada;

Hargreaves, A. & Shirley, D. (2012). *The global fourth way: The quest for educational excellence*. Thousand Oaks, CA: Corwin.

88 EQAO stands for Education Quality and Accountability Office. It is an independent agency of the Ontario provincial government, with a chief responsibility for preparing, administering and reporting province-wide standardized assessments in selected grades and subjects.

89 Fullan, M. (2013). *Great to excellent: Launching the next stage of Ontario's education agenda*. Retrieved May 31, 2013, from http://www.michaelfullan.com.

90 Campbell, C. & Fulford, D. (2009). *From knowledge generation to knowledge integration: Analysis of how a government uses research*. Paper presented at 2009 AERA Annual Meeting. Retrieved October 19, 2013, from http://www.edu.gov.on.ca/eng/research/AERA2009_KIPaper.pdf.

91 Fuhrman, S. & Elmore, R.F. (Eds.) (2004). *Redesigning accountability systems for education*. New York: Teachers College Press;

Sharratt, L. & Fullan, M. (2012). *Putting FACES on the data: What great leaders do!* Thousand Oaks, CA: Corwin;

Earl, L.M. & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement*. Thousand Oaks, CA: Corwin.

92 Glaze, A. & Campbell, C. (2007). Putting literacy and numeracy first: Using research and evidence to support improved student achievement. Paper presented at 2007AERA Annual Meeting. http://www.edu.gov.on.ca/eng/research/litNumfirst.pdf.

93 Heritage, M. & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. *Yearbook of the National Society for the Study of Education*, *104*(2), 320-339;

Campbell, C. & Fulford, D. (2009). *From knowledge generation to knowledge integration: Analysis of how a government uses research*. Paper presented at 2009 AERA Annual Meeting. Retrieved October 19, 2013, from http://www.edu.gov.on.ca/eng/research/AERA2009_KIPaper.pdf.

94 Glaze, A. & Campbell, C. (2007). Putting literacy and numeracy first: Using research and evidence to support improved student achievement. Paper presented at 2007AERA Annual Meeting. Retrieved October 19, 2013, from http://www.edu.gov.on.ca/eng/research/litNumfirst.pdf.

95 See also Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data*. Los Angeles: University of Southern California, Center on Educational Governance.

96 Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data*. Los Angeles: University of Southern California, Center on Educational Governance;

Sharratt, L. & Fullan, M. (2012). *Putting FACES on the data: What great leaders do!*. Thousand Oaks, CA: Corwin;

Hargreaves, A. & Braun, H. (2012). *Leading for All: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

97 Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231-268;

See also Hargreaves, A. & Fink, D. (2012). *Sustainable leadership*. San Francisco, CA: Jossey-Bass.

98 Bird, S., Cox, D., Farewell, V., Goldstein, H., Holt, T. & Smith, P. (2005). Performance indicators: good, bad and ugly. *Journal of the Royal Statistical Society: Series A*, 168, Part 1;

Daly, A. (2009). Rigid response in an age of accountability: The potential of leadership and trust, *Educational Administration Quarterly*, *45*(2), 168-216.

99 Little, J. (1990). The persistence of privacy: Autonomy and initiative in teachers' professional relations. *Teachers College Record*, *91*(4), 509-536;

Rosenholtz, S.J. (1989). *Teachers' workplace: The social organization of schools*. New York: Longman;

Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data*. Los Angeles: University of Southern California, Center on Educational Governance.

100 The Organisation for Economic Co-operation and Development (2012). *Equity and quality in education: Supporting disadvantaged students and schools.* Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org.

101 Sahlberg, P. (2011). *Finnish lessons*. New York: Teachers College Press.

102 The Organisation for Economic Co-operation and Development (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org;

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. New York: McKinsey & Company.

103 Sharratt, L. & Fullan, M. (2009). *Realization: The change imperative for deepening district-wide reform*. Thousand Oaks, CA: Corwin.

104 Wilkinson, R.G. & Pickett, K. (2010). *The spirit level: Why greater equality makes societies stronger*. New York: Bloomsbury Press;

UNICEF Office of Research (2013). 'Child well-being in rich countries: A comparative overview.' *Innocenti Report Card 11*. Florence: UNICEF Office of Research.

105 Hatch, T. (2001). It takes capacity to build capacity, *Education Week*, *20*(22), 44-47.

106 Bryk, A., Sebring, P., Allensworth, E., Luppescu, S., & Easton, J.Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

107 Payne, C.M. (2008). *So much reform, so little change: The persistence of failure in urban schools*. Cambridge, MA: Harvard Education Press.

108 Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data*. Los Angeles: University of Southern California, Center on Educational Governance.

109 Datnow, A. (2011). Collaboration and contrived collegiality: Revisiting Hargreaves in the age of accountability. *Journal of Educational Change*, *12*(2), 147-158.

110 Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data*. Los Angeles: University of Southern California, Center on Educational Governance.

111 Mintrop, H. (2004). *Schools on probation*. New York: Teachers College Press;

Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, *42*(2), 231-268;

Baker, M. & Foote, M. (2006). Changing spaces: Urban school interrelationships and the impact of standards-based reform. *Educational Administration Quarterly*, *42*(1), 90-123;

Hargreaves, A. (2003). *Teaching in the knowledge society: Education in the age of insecurity*. New York: Teachers College Press;

Falk, J. & Drayton, B. (2004). State testing and inquiry-based science: Are they complementary or competing reforms? *Journal of Educational Change*, *5*(4), 345-387.

112 Kearns, D.T. (1990). Leadership through quality. *The Executive*, *4*(2), 86-89;

Tucker, M.S. (2009). Industrial benchmarking: A research method for education. In A. Hargreaves & M. Fullan (Eds.), *Change Wars*. Bloomington, IN: Solution Tree.

113 The Organisation for Economic Co-operation and Development (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org;

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company. Retrieved May 30, 2013, from http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools/ ;

Tucker, M.S. (2011). *Standing on the shoulders of giants: An American agenda for education reform*. Washington, DC: National Center on Education and the Economy.

114 The Organisation for Economic Co-operation and Development (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org/.

115 See also Hargreaves, A. & Shirley, D. (2012) *The global fourth way: The quest for educational excellence*. Thousand Oaks, CA: Corwin.

116 Hargreaves, A., Shirley, D., Evans, M., Stone-Johnson, C., & Riseman, D. (2006). *The long and the short of school improvement: Summary of the evaluation report on the Raising Achievement Transforming Learning Project of the Specialist Schools and Academies Trust*. London: SSAT;

Hopkins, D. (2009). Every school a great school: Realising the potential of system leadership. In A. Hargreaves, A. Lieberman, M. Fullan, & D. Hopkins (Eds.), *Second International Handbook of Educational Change* (741-764). Dordrecht, Netherlands: Springer.

117 Reason Foundation (2013). *Weighted student formula in the states.* Retrieved May 26, 2013, from http://reason.org/news/show/apr-2013-weighted-student-formula/.

118 Hargreaves, A. & Shirley, D. (2012) *The global fourth way: The quest for educational excellence*. Thousand Oaks, CA: Corwin.

119 Gates, B. (2013, January 25). Bill Gates: My plan to fix the world's biggest problems. *The Wall Street Journal.* Retrieved May 31, 2013, from http://online.wsj.com/article/SB10001424127887323539804578261780648285770.html;

Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey & Company. Retrieved May 30, 2013, from http://www.mckinsey.com/client_service/social_sector/latest_thinking/worlds_most_improved_schools/.

120 Glaze, A. & Campbell, C. (2007). Putting literacy and numeracy first: Using research and evidence to support improved student achievement. Paper presented at 2007AERA Annual Meeting. Retrieved October 19, 2013, from http://www.edu.gov.on.ca/eng/research/litNumfirst.pdf;

Hargreaves, A. & Fullan, M. (2012). *Professional capital: Transforming teaching in every school*. New York: Teachers College Press.

121 Hargreaves, A. (2003). *Teaching in the knowledge society: Education in the age of insecurity*. New York: Teachers College Press;

Falk, J. & Drayton, B. (2004). State testing and inquiry-based science: Are they complementary or competing reforms? *Journal of Educational Change*, *5*(4), 345-387.

122 Jesson, D., Mayston, D., & Smith, P. (1987). Performance assessment in the education sector: Educational and economic perspectives. *Oxford Review of Education*, *13*(3), 249-266;

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal*, *27*(4), 433-442.

123 Duncan, A. (2011, September 23). [Letter to Chief State School Officers]. U.S. Department of Education. Retrieved May 30, 2013, from http://www2.ed.gov/policy/gen/guid/secletter/110923.html.

124 Braun, H.I. (2005). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.). Value-added models in education: Theory and applications (19-38). Maple Grove, MN: Jam Press;

Linn, R.L., Baker, E.L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3-16.

125 Hout, M. & Elliott, S.W. (Eds.). (2011*). Incentives and test-based accountability in education.* Washington, DC: National Academies Press;

Braun, H.I. (2005). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.). Value-added models in education: Theory and applications (19-38). Maple Grove, MN: Jam Press;

Hargreaves, A. & Harris, A. (2011). *Performance beyond expectations.* Nottingham, UK: National College for School Leadership.

126 Reardon, S.F. & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492-519;

Baker, E.L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Rothstein, R., Ravitch, D., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper No. 278). Washington, DC: Economic Policy Institute;

Braun, H.I. (2005). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.). Value-added models in education: Theory and applications (19-38). Maple Grove, MN: Jam Press.

127 McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND;

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23);

Baker, B., Oluwole, J., & Green III, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Education Policy Analysis Archives, 21*(5).

128 Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23).

129 Fullan, M., Hill, P., & Crévola, C. (Eds.). (2006). *Breakthrough*. Thousand Oaks, CA: Corwin.

130 Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11.

131 Koretz, D.M. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

132 Elmore, R.F. (2004). Conclusion: The problem of stakes in performance-based accountability systems. In S. Fuhrman, S. & R.F. Elmore (Eds.) *Redesigning accountability systems for education* (274-296). New York: Teachers College Press.

133 Hout, M. & Elliott, S.W. (Eds.). (2011*). Incentives and test-based accountability in education.* Washington, DC: National Academies Press;

Rothstein, R., Jacobsen, R., & Wilder, T. (2008). *Grading education: Getting accountability right*. Washington, DC: Economic Policy Institute;

Springer, M.G. (2009). *Performance incentives" Their growing impact on American K-12 education*. Washington, DC: Brookings Institution Press;

Pink, D.H. (2009). *Drive: The surprising truth about what motivates us*. New York: Riverhead Books.

134 Hargreaves, A. & Shirley, D. (2012). *The global fourth way: The quest for educational excellence*. Thousand Oaks, CA: Corwin;

Tucker, M.S. (2011). *Standing on the shoulders of giants: An American agenda for education reform*. Washington, DC: National Center on Education and the Economy;

The Organisation for Economic Co-operation and Development (2011). *Strong performers and successful reformers in education: Lessons from PISA for the United States*. Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org;

Sahlberg, P. (2011). *Finnish lessons*. New York: Teachers College Press.

135 Daly, A.J., Der-Martirosian, C., Ong-Dean, C., Park, V., & Wishard-Guerra, A. (2011). Leading under sanction: Principals' perceptions of threat rigidity, efficacy, and leadership in underperforming schools. *Leadership and Policy in Schools*, *10*(2), 171-206.

136 Marsh, J.A., Pane, J.F., & Hamilton, L.S. (2006). *Making sense of data-driven decision making in education evidence from recent RAND research*. Santa Monica, CA: RAND.

137 Datnow, A., Park, V., & Wohlstetter, P. (2007). *Achieving with data.* Los Angeles: University of Southern California, Center on Educational Governance.

138 Wood, D. (2007). Teachers' learning communities: Catalyst for change or a new infrastructure for the status quo? *Teachers College Record*, *109*(3), 699-739;

Hargreaves, A. & Braun, H. (2012). *Leading for all: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

139 Hargreaves, A. & Braun, H. (2012). *Leading for All: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

140 Hargreaves, A. & Braun, H. (2012). *Leading for All: A research report of the development, design, implementation and impact of Ontario's "Essential for Some, Good for All" initiative*. Ontario: Council of Ontario Directors of Education.

141 Hargreaves, A. & Shirley, D. (2009). *The fourth way: The inspiring future for educational change*. Thousand Oaks, CA: Corwin.

142 Sahlberg, P. (2011). *Finnish lessons*. New York: Teachers College Press;

The Organisation for Economic Co-operation and Development (2012). *Equity and quality in education: Supporting disadvantaged students and schools.* Paris: Author. Retrieved June 2, 2013, from http://www.oecd-ilibrary.org.