Governing Public School Districts: Insights from Economics, Political Science, and Business Management

by

Fred Thompson
Director, Center for Governance and Public Policy Research

Polly Rizova
Associate Professor of Management

Henry Bi
Associate Professor of Information Systems

Atkinson Graduate School of Management
Willamette University
900 State Street
Salem OR, 97301 USA

February 28, 2011

Abstract

In this essay, we review the contributions economists, political scientists, and business management scientists have offered to an understanding of how local school districts, boards of education, school superintendents, and district offices affect student learning and other achievement outcomes of interest; where and how their work differs from that done by students of educational administration; and the implications of their findings for future research.

Keywords: School districts, Performance measurement, Governance, Field experiments
Governing School Districts: Insights from Economics, Political Science, and Business Management

What do we know about the frontline institutions that govern America’s public schools – local school districts, boards of education, school superintendents, and district offices? According to Deborah Land (2002), not much. Recognizing this state of affairs, the Institute for Education Sciences made the organization and management of school districts one of its research emphases for 2010-2011.

This integrative review was occasioned by the observation that, if we are at or near the starting line, it might be interesting to see what other discourse communities concerned with the governance of large-scale organizations – economics, political science, and management (which we have taken to comprehend organizational sociology) – have to say, if anything, about the organization and management of school districts. To be sure, educational researchers concerned with district governance (Bowers, 2010; Elmore, 1999-2000; Honig, 2008, 2009; York-Barr and Duke, 2004) often cite bits and pieces from these literatures, but often, it seems to us, unsystematically.

Were this impression correct, it would hardly be surprising. School districts rarely attract the attention of economists, political scientists, and management scientists because they find them intrinsically interesting, but because they are convenient vehicles for addressing the theoretical and methodological concerns that are idiosyncratic to their own disciplines. As Meier and Krause (2006: 301), for example, explain, school districts are ideal subjects for anyone who wants to study bureaucratic behavior. “Most bureaucracies are local. If scholars want a situation in which one can turn some of our variables into constants and get good comparability across agencies, then the best strategy is to select
the same type of agency across states or local jurisdictions. The most common bureaucracies in both number and employment are schools.” Moreover, economists, political scientists, and management scholars generally presume that student achievement metrics make good outcome measures. Consequently, their research on school districts is usually only indirectly concerned with matters of interest to the educational research community. Even, in the minority of cases, where their research focuses on the effect of district governance on teaching and learning, matters of direct concern to educational researchers, it is rarely addressed to them, but must be looked for.

This review focuses on the approaches taken by of economists, political scientists, and management scholars concerning the effects of school-district design and context features on teaching practices and learning outcomes and their findings. Ultimately, we seek to understand how certain core projects carried out by district leadership are executed. These are: designing system improvement strategies, designing and executing incentive structures for schools, principals, and teachers, recruiting and evaluating principals, providing professional development, allocating system resources, and buffering principals, teachers, and students from extraneous and distracting non-instructional issues. We are also interested in finding out the amount of time and energy devoted to these matters and the ways in which applying different approaches to executing these projects as well as the contexts in which they occur, affect success or failure of school districts.

Hence, our specific interest is in school districts and their management. Local school districts are creatures of state governments, with their purposes, powers and jurisdictions, and governance arrangements specified in state statutes. Therefore, states
(and increasingly the federal government) play an important role in the governance of schools.\(^1\) Arguably, the rise in state and federal authority over schools at the expense of local district autonomy is one of the main education governance themes of the last 60 years.\(^2\); the examination of its role though, is out of the scope of this essay. Other governance issues omitted from this survey include the role of the courts and the effects of court decisions on district governance and the reform strategies open to them, state policies with respect to teachers’ unions, except where mediated through differential effects on district-level governance, the effects of competition and privatization, and mayoral control (see Hill 2003, 2004).

Most school districts have been set up as independent, special-purpose tax districts, run by elected boards. Consequently, they vary on a number of dimensions, some of which are malleable (under the immediate control of districts), many however, are not. For example, districts can influence the number and size of their schools, staffing, salaries, benefits, and incentive and evaluation systems, budget policies and practices and resource allocations, procurement policies, as well as curricula, and educational doctrine and aims. At the same time, they have little influence over the size or scope of their jurisdictions, the number of students they oversee, the backgrounds, characteristics, mobility, or diversity of their students, the variety or complexity of their

---

\(^1\)See Brewer and Smith (2006) and Smith, Ahn, and Brewer (2007) for an overview of school governance.

\(^2\)We would note that economists (e.g. Schoderbek and Mensah, 2010) and political scientists (e.g. O’Toole and Meier 2011) have consistently found that student achievement varies directly with the proportion of funds raised locally, which is consistent with the notion that local property-tax funding (as opposed to centralized financing) promotes efficiency in local public-good provision (Hoxby, 1999; Fischel, 2001, 2009). It is equally consistent with the notion that attention to demands of multiple external stakeholders comes at the expense of attention to core projects.
funding sources, total budgets or per-pupil spending, or union power. We treat malleable dimensions of the district environment as design features in the analysis which follows; non-malleable factors as context features.

School districts are different from other kinds of local jurisdictions in at least two ways. Most states require school-board elections to be non-partisan and require school boards to appoint professional administrators, superintendents, as their chief operating officers. These mandates reflect the unique importance of education and the progressive-era belief that this purpose would be better served if school districts were “insulated” from politics and run by qualified professionals, such as district superintendents.

District superintendents typically have a great deal of discretion: they set agendas for school-board meetings, propose district budgets, establish schools’ curricula, and oversee personnel processes. Superintendents are generally vested with authority to hire and fire managers (principals, assistant superintendents, etc.) and general authority to

---

3 As O’Toole and Meier explain (2011:341), “the use of the special-district form in this sector was motivated by the objective of insulating public education from politics as usual and from dependence on or competition with other forms of public services.” Nevertheless, an Organization for Economic Cooperation and Development (Meuret, Prodhom, and Stocker 1995: 52) study of 14 national education systems found that more education decisions must made in consultation with others (44 percent) in the U.S. than in any other Western democracy. U.S. systems were also rated the lowest in terms of local (that is, school district) autonomy. That observation aside, this review generally omits from consideration analyses of the causes of productivity heterogeneity based on international and inter-state comparisons using PISA, TIMSS and NAEP scores, primarily because they do not address district-level governance issues, but also because differences seem to reflect deep structural features at the societal level, although it looks like there is something special about Finland and its school system (Andersen, 2010).

4 Indeed, based on structured interviews with 490 school board members and 82 school superintendents and briefer interviews with 91 board members and 12 superintendents from districts in 13 large metropolitan areas, Zeigler, Jennings, and Peak (1974) found that superintendents generally dominated district boards, especially in districts with relatively politically responsive school boards. Unfortunately, this study has not been replicated, so we cannot say whether superintendent dominance has altered over time.
move personnel to different locations or positions. In the typical school district the superintendent’s staff or central office is responsible for financial management, human-resource management, purchasing, data gathering, and system leadership. In other words, it is the role and prerogative of the office of the superintendent to “design” the core projects in its district in a manner such that it is conducive to the learning outcomes it seeks to achieve. The examinations of how school districts, namely their leadership, exercise these responsibilities and the effect of those on student learning is at the center of this paper.

*The District Office as a Black Box: Insights from Economics*

As a general rule, economists focus on cost and production behavior and their drivers. They think in terms of inputs, outputs, and constraints. In other words, their view of school districts looks something like this:

\[
\text{District outputs or learning outcomes} = f(\text{District inputs, student characteristics})
\]

Obviously, this view raises questions about measuring outputs (what goes into the response matrix?), functional forms to be estimated (what goes into the predictors matrix, according to what specification?), and how best to estimate these relationships (statistically or computationally and with which techniques, regression or multilevel/hierarchical models?). It also treats the conversion process as a black box: inputs in, outputs out. As a consequence, economists, with a few salient exceptions, tend to treat the malleability or controllability of district or school-level variables as a distinctly secondary issue.

that school districts employ $m$ inputs $X = (x_1, \ldots, x_m)$ – inputs are things that must be purchased such as labor (teachers, etc.) and capital (plant and equipment, etc.) – to generate a vector of services $Y = (y_1, \ldots, y_n)$ or final outcomes (test scores, the percentage of students going to higher education, etc.), subject to the influence of a set of $k$ constraints $Z = (z_{j1}, \ldots, z_{jk})$ or environmental variables – givens that are not purchased (student characteristics, etc.) that influence the transformation of district inputs into district outputs. Specifically, Keisling modeled district production of services as a simple additive function of various combinations of input levels, holding environmental variables constant:

$$ (Y) = f(X|Z). \quad (1) $$

He estimated these relationships using New York school district data with ordinary least squares regression. He found no relationship between inputs and outputs: per pupil spending, teachers per pupil, plant and equipment per pupil, teacher experience, and teacher degree level do not predict student achievement or their effect is inconsistent across alternative specifications. He also found that performance in small school districts (fewer than 2,000 pupils in average daily attendance – ADA) was highly unstable and that economies of scale with respect to expenditures per student exist up to an ADA of about 3000 students, but as school districts become larger, test scores decline. Subsequent analysis has generally confirmed Kiesling’s findings.\(^5\)

While many economists have continued to pursue Kiesing’s approach – building bigger and better data sets, improving outcome and input measures and controls, using

\(^5\)On the indeterminacy of spending, see Hanushek (2009), inputs, Hanushek (2003) and teacher characteristics, Wayne and Youngs (2003); on noise to signal ratios, see Kane and Staiger (2002) and Staiger and Rockoff (2010) and scale, Jacques, Brorsen, and Richter (2000).
more sophisticated econometric methods, others have concluded we should, instead, focus on finding ways to improve educational governance. Their efforts have taken two distinct paths. One path embraces Kiesing’s methods, but seeks to understand the relationship between district routines\(^6\) and student achievement. For the most part, economists who take this path have focused on a fairly narrow set of practices: how principals and teachers are selected and retained, how they are rewarded and controlled; how their performance is assessed and how those assessments are used. Nevertheless, one could generalize this path as modeling district production of services as a function of an array of organizational features \(O = (o_1, ..., o_j)\), holding environmental variables and inputs constant:

\[
(Y) = g(O|Z,X).
\]

The other path rejects Kiesling’s methods. Its followers argue that what we want to do is replace the existing state of affairs with a better one, that the gap between ‘what is’ and ‘what might be’ can be bridged best by identifying outstanding performers and explaining their performance, and that we cannot get there by OLS or, at least, not by

\(^6\)In economics, an organization is a collection of people, an architecture, a set of routines, and a culture (Roberts, 2004: 17). By architecture economists mean the division of authority and responsibility within the organization and the incentive and coordination mechanisms it employs; by routines they mean its processes, policies, and procedures; by culture they mean the mental models of its members and the “norms of behavior that prevail in dealing with other members of the [organization] and outsiders” (18). Because economists are more comfortable talking about incentives than about coordination, they tend to give coordination short shrift, culture too. Nevertheless, there is now an array of models, applying the optimization and equilibrium concepts of economics, which account for the heterogeneity of design features found in organizations. Unfortunately, the evidentiary basis for organizational economics rests largely, although by no means entirely, on case studies and anecdote rather than solid data. Moreover, it is not clear whether design features are primarily embodied in people or organizational culture and are, therefore, fundamentally nontransferable or if they can be transferred between organizations like any other technology.
OLS alone, because regression methods tend to downplay the unusual and focus on the norm or the average.

**Discrete Practices: Incentives, Recruitment/Retention**

What causes differences in organizational productivity? Is governance or management crucial to these differences? While researchers in business schools and consultancies typically make strong claims about the effects of management on productivity, economists are often skeptical, noting that the claims made by researchers in business schools and consultancies are typically based on surveys and case studies and that both forms of evidence are problematic for making causal inferences. Economists generally take the view that in order to answer these questions one must do three things: quantify management practices/design features and productivity measures across different organizations in a comparable way (see Bloom and Van Reenen, 2007 and 2010 for discussions), show that the correlation between design features and productivity is causal, and provide mechanism-based theories that account for relationships between design features and productivity (Pawson and Tilley, 1997). Further, they argue that field

---

We do not necessarily share this skepticism. For example, there is a growing literature looking at the various “hard” technological variables that influence productivity – especially information and communication technologies, ICT, which shows that the effect of technology on productivity depends on how organizations are managed. Bloom, Sadun and Van Reenen (2007) for example, find that the effects of investment in ICT are stronger for organizations that base hiring, pay and promotion decisions on effort/ability and not just tenure, employ rigorous procedures for dealing with underperformers, etc. Brynjolfsson and Hitt (2000, see also Roberts, 2004 and Thompson, 2006) agree, but emphasize the importance of adopting digital processes and distributing decision rights to frontline personnel. They claim that these are the design features that distinguish effective ‘netcentric’ organizations from traditional bureaucracies.
trials/experiments are the one best way to confirm the causal effects of management practices on productivity.

With respect to school districts, there is some evidence that governance matters. Based on an analysis of fixed effects, Iatarola, Schwartz, and Stiefel (2002) find that districts influence school performance. They speculate that teacher recruitment and retention practices are the primary determinants of these results, but do not test their speculations. Schoderbek and Mensah (2010) argue that superintendent salaries are an effective proxy for managerial quality and find that salaries are positively correlated with test scores (although they report that administrative costs excluding salaries are negatively correlated with test scores, see also Brewer 1996\(^8\)). Of course, this could be due to reverse causation. More productive school districts are, perhaps, more likely to retain superintendents and, if wages are linked to seniority, pay them more. However, in this case we can report that O’Toole and Meier (2011) obtain similar results and find that superintendent salaries Granger-cause educational productivity (see below). To an economist, one of the more persuasive warrants for the belief that districts make a positive contribution to educational productivity lies in the fact that independent and charter schools increasingly and voluntarily choose to organize themselves into multi-school systems (Miron and Urschel, 2010; Hannaway and Kimball, 1998).

This begs the question “why?” Can the causal mechanisms that would account for such an outcome be identified reliably? To answer this question, economists tend to

---

\(^8\)Using 1978–87 panel data from school districts in New York State, Brewer looks for a systematic relationship between administrative spending and educational productivity in the form of standardized test scores and obtains inconsistent results from a variety of statistical models.
look to policies that link hiring, retention, pay and promotion decisions to productivity.\footnote{Another preoccupation of economists is size: district, school, and class. Thanks to the Tennessee class-size experiments, it now seems clear that class size matters: that smaller is better, although how much is still up for grabs. What the Tennessee studies most clearly demonstrate is the analytic payoff to randomized trials. In these studies, both students \textit{and} teachers were randomly assigned to treatment and control groups. It now seems likely that the previous conventional wisdom, which held that class size didn’t matter, was confounded by two factors: other things being equal, reducing class size dilutes teacher quality and the rational propensity of principals to assign better teachers to larger classes, just as the still earlier conventional wisdom, that it did, was confounded by a correlation between class size and personal, home, and community differences that affected students’ test scores. On school size, see Leithwood and Jantzi (2009).} Mostly, they have focused on policies that affect teachers rather than principals or other school employees. As Hanushek (2010: 1) states, this is a smart place to start because “high quality teachers are the most important assets of schools.” He explains:

A teacher one standard deviation above the mean effectiveness annually generates marginal gains of over $400,000 in present value of student future earnings with a class size of 20 and proportionately higher with larger class sizes. Alternatively, replacing the bottom 5-8 percent of teachers with average teachers could move the U.S. near the top of international math and science rankings with a present value of $100 trillion.

He however, takes the view that it is impossible to identify good teachers before they get into the classroom and, even then, it is not easy. Wayne and Youngs (2003), on the other hand, note that the selectivity and reputation of the colleges that teachers attended and teachers’ test scores are correlated with student learning and that high-school mathematics students learn more from teachers who have studied mathematics than from those who haven’t, but they admit that most of these relationships are not very strong and our knowledge is inconclusive.
Incentives. Some economists believe the main problem with schools is a lack of extrinsic incentives: if school personnel were rewarded based on student performance, school productivity would improve. A compelling study, which makes this case, is the one by Muralidharan and Sundararaman (2009). The authors provide evidence that school-level group incentives and a fortiori teacher-level individual incentives increase learning. Their study is especially convincing for two reasons: their results are drawn from data from a very large sample, 500 schools covering approximately 55,000 students, and a carefully stratified experimental design in which schools were randomly assigned to treatment and control groups. The primary mechanism through which incentives improve learning appears to be more intensive teaching and attention to weaker students. Muralidharan and Sundararaman’s (2009) venue, however, was India. One might, therefore, question their study’s direct relevance to district policy making in the United States, although we are inclined to think that it might very well offer useful advice to districts with similar academic achievement levels (poor to fair) and an equally ambitious corps of teachers.

An experiment conducted over a three-year period (2006-07 through 2008-09) in the Metropolitan Nashville School System by Springer and his colleagues (Springer, Ballou, Hamilton, Le, Lockwood, McCaffrey, Pepper, and Stecher 2010) however, decisively rejects the hypothesis that, if teachers were rewarded based on student performance, their teaching effectiveness would improve. In this study, middle-school mathematics teachers voluntarily agreed to participate in a controlled experiment designed to assess the effects of financial rewards. The volunteers were then randomly assigned to a treatment group (eligible for significant bonuses if their students showed large gains on standardized tests).
and a control group (not eligible for bonuses). The findings demonstrate that students in the treatment group did not outperform the control group. The authors explain:

The experiment was intended to test the notion that rewarding teachers for improved scores would cause scores to rise. It was up to participating teachers to decide what, if anything, they needed to do to raise student performance: participate in more professional development, seek coaching, collaborate with other teachers, or simply reflect on their practices. Thus, the Project on Incentives in Teaching was focused on the notion that a significant problem in American education is the absence of appropriate incentives, and that correcting the incentive structure would, in and of itself, constitute an effective intervention that improved student outcomes (Springer et al. 2010: ix).

To a degree, economists seem to treat teachers as interchangeable, but largely autonomous units. The implicit assumption is that they know how to do their job. If they aren’t sufficiently productive, the problem is one of extrinsic motivation. At a minimum, Springer, et al. (2010) findings should lead us to question that assumption. Instead, it probably makes sense to suppose that teachers, like other personnel, are more productive when they are well integrated into a community of practice, engaged in an ongoing dialogue about their collective enterprise, and learn from each other (see Brown and Duguid, 1991; Leana and Pil, 2006; Pil and Leana, 2009). This is a standard insight from the economics of development. If, for example, there are ten members in a group and each figures out how to be ten percent more productive, the group as a whole is ten percent more productive. But, if they share what they have learned, even if only half of what they learn is transferable, the group will be 63 percent more productive (of course,
if they transfer all their knowledge, that produces an even better result, a productivity increase of 160 percent).

Communities of Practice. We see two problems with testing this presumption; both have to do with assessing treatment effects. First, it is difficult to quantify complex management practices/design features across different organizations in a comparable way. This means that, even if identification is not defeated by multiple, contingent causation (see below), errors in variables and/or multicollinearity can severely bias OLS results. Consequently, this option rarely works satisfactorily where complex organizational interventions are concerned. Well-designed experiments obviate the need to quantify complex management practices/design features across different organizations in a comparable way (as in the experiments described immediately above), but experiments testing interventions require large numbers of subjects (or huge treatment effects) to produce conclusive results. Organizations make awkward subjects, which gives rise to our second problem, insuring that all of the organizations in the treatment group carry out the same intervention, while at the same time insuring that controls do not emulate the design features adopted by treatment group. This means that the design features of the intervention must be meticulously codified and careful attention given to its execution to insure that each and every treatment site follows the code laid out for it and to insure that control sites are not influenced by treatment sites.

One prevalent field test based on the presumption that building an effective community of educational practice is the key to improving the productivity of schools is the Teacher Advancement Program (TAP). Lowell Milken and colleagues developed this program at the Milken Family Foundation in the 2000-2001 school year. The National
Institute for Excellence in Teaching (NIET) now manages it and supervises its execution at treatment sites. TAP has been adopted by nearly 100 school districts across the country. It now involves over 200 schools, 7,500 teachers and 85,000 students (Daley and Kim, 2010). For our purposes, the most important facts about TAP are the codification of its design features and the care with which NIET monitors and enforces execution of the program.

TAP comprehends five key elements or features (Daley and Kim, 2010). First, to participate in TAP, a school’s faculty must voluntarily agree to do so and funds must be raised to finance a bonus pool, which usually means that the district must find a local, state or NIET sponsor. Second, NIET requires the use of a standardized instructionally focused, teacher-evaluation system. This system is designed to support both collective learning and accountability and includes a qualitative component, which relies on classroom observation by multiple trained raters using a research-based rubric, and a quantitative component, which measures the teacher’s and the school’s value-added contributions to the achievement gains of students. ¹⁰ Third, NIET trains and certifies a cadre of instructional leaders, master and mentor teachers, who conduct classroom observations, lead their colleagues in weekly cluster-group meetings, and, together with principals, build and sustain effective communities of practice. Fourth, led by master and mentor teachers, TAP teachers participate in weekly cluster-group meetings, where they

¹⁰Value-added assessment involves matching each student’s test scores to his or her own previous scores, which reflect differences in background factors that influence student attainment, and measuring the student’s academic growth as the change in attainment from the beginning of some period to the end, in this case from the beginning to the end of the academic year. According to Klitgaard and Hall (1973) the value-added assessment and gain approaches to measuring institutional effectiveness were independently formulated by Henry S. Dyer (1972) and Stephen Barro (1970).
examine student data, engage in collaborative planning, and learn from each other’s experiences. And, fifth, to focus their attention, sustain commitment to collaborative planning and problem solving, and, perhaps, motivate them. Teachers in TAP schools “earn bonuses … based on their observed skills, knowledge, and responsibilities; their students’ average growth in achievement; and the entire school’s average growth in achievement. Master and mentor teachers receive additional compensation based on their added roles and responsibilities. Combining these sources, performance pay for a teacher in a TAP school can reach up to $20,000” (Daley and Kim, 2010: 9).

There have been four assessments of TAP, two done in house by NIET (Solmon, White, Cohen and Woo, 2007; Daley and Kim, 2010) and two done by independent analysts (Springer, Ballou and Peng, 2008; Glazerman and Seifullah, 2010). Unfortunately, TAP’s requirement of a sponsor militates against the use of a true experimental design to assess its effectiveness. Consequently, all four assessments adopt similar methodologies, so-called quasi-experimental designs, in which TAP schools were matched to non-TAP schools whose student and institutional characteristics were as nearly similar to the TAP schools as possible.¹¹

There were two main differences in study designs (aside from the independence of the second set). First, the NIET researchers’ treatment groups comprehended all TAP schools, while their control groups were made up of a large sample of non-TAP schools.

¹¹For example, Glazerman and Seifullah (2010) used a logistic regression model, in which school size, teacher retention, the number of years since the school met the state’s standards for adequate yearly progress, average math and reading scores on the Illinois Standards Achievement Test (standardized within grade across the district Student race/ethnicity), student poverty, and truancy rates were entered into the predictors matrix, to estimate the probability that a school would be selected into the TAP finalist pool. They then used the propensity scores assigned to each school in the district by this model to match non-TAP schools to TAP schools.
from outside TAP districts; the independent researchers’ treatment groups included subsets of TAP schools (all TAP schools in two states: Springer et al. 2008; TAP schools in the Chicago school district: Glazerman and Seifullah, 2010), while their controls were made of matched pairs of non-TAP schools from the same districts.

Second, the NIET researchers used the same value-added scores (students’ current scores on annual standardized assessments controlled for their previous scores) to assess treatment effects that they used to evaluate the achievement gains of students under TAP. The independent researchers used statistically adjusted (using student characteristics) current scores on annual standardized assessments to compare the productivity of TAP to non-TAP schools.¹²

Solmon et al. (2007) report that TAP schools stochastically dominate matched non-TAP schools in terms of student achievement growth (see Figure 1). Daley and Kim (2010) report some of the mechanisms that would produce such an outcome: TAP schools do a better job of retaining skilled teachers, the skill measures they use are positively and significantly correlated with student achievement growth, and TAP teachers increase in observed skill levels over time. In contrast, Springer et al. (2008) found positive effects for TAP at elementary grades but undetectable or negative impacts at middle and high school grades. Moreover, they report evidence of selection effects, which is not surprising given the screening and self-selection that must take place for a

¹²Springer et.al (2008) used panel-data sets of math scores. Glazerman and Seifullah (2010) used statistically adjusted math and reading scores on the Illinois Standards Achievement Test, but also replicated their analysis using value-added scores. They got practically identical results using the two different measures in their response matrices, but claim that using statistically adjusted average test scores is more reliable than using value-added scores. Because the results are the same, this claim doesn’t really matter, but given the small size of their data set and the reduction in degrees of freedom entailed by their preferred procedure, it is not obvious why it should be valid.
school to adopt TAP. Finally, Glazerman and Seifullah (2010) found no evidence whatsoever that participation in TAP is associated with increases in student test scores. They further report that their evidence is clearly inconsistent with “a hypothesis that TAP has a more beneficial effect as schools spend more time with the program.”

Figure 1: TAP Schools versus Control Schools Cumulative Distribution Comparative Plot

Standardized School Effectiveness Estimates

Springer et al. (2008) report that students in both TAP and non-TAP school made remarkable achievement gains; Glazerman and Seifullah (2010) report that students in neither the TAP schools in their sample nor the matched non-TAP schools made significant achievement gains during the two years studied.

We don’t know what to make of these results. Methodologically, we like the in-house studies better than the independent studies. Solmon and his colleagues show that
that a well-defined set of design features are correlated with educational outcomes and provide good reasons to believe the direction of causation goes from design features to outcomes rather than the other way around. Daley and Kim (2010) provide mechanism-based theories that account for this relationship and persuasive evidence those theories make sense. At the same time we also appreciate the subtle and often unconscious biases that influence researchers’ findings. One cannot simply reject the possibility the results reported by Solmon and his colleagues reflect wishful thinking on their part (Borman, Hewes, Overman and Brown, 2003). The one inference we are inclined to draw is that TAP may not be entirely fitting to the schools in Chicago school district that have adopted it. Its design features seem more appropriate to the problem of moving a district from fair to good or good to great than for moving from poor to fair.

Searching for Unusually Effective Districts

Best-practice analysis is the other path frequently taken by productivity researchers. Its logic is straightforward: find the best performers in a particular field (benchmark), figure out how their practices differ from representative performers, and copy those differences. The classic example of this kind of research is Womack, Jones, and Roos’ The Machine that Changed the World (1990), which looked at manufacturing productivity in the automobile industry and identified Toyota as the world’s benchmark.14

13 In this vein, see also Borman and Dowling (2008) and York-Barr and Duke (2003).
14 The Toyota Production System, implemented by Eiji Toyoda and Taiichi Ohno, is usually identified with certain design features: total quality management, quality circles, lean manufacturing, just-in-time delivery, and price based costing. Womack et al. (1990) propose that these design features are, instead, consequences of its reliance on worker learning and sharing. Toyota emphasizes that nobody in the production process but the frontline worker adds value, that the frontline worker can, given the means, motive, and
However, it is a little known fact that best-practice analysis was inspired by three economists, Michael J. Farrell, Bob Klitgaard and George Hall. Farrell was concerned about productivity in general; Klitgaard and Hall were concerned specifically with educational productivity.

Farrell (1957) and Farrell and Fieldhouse (1962) observed that productive efficiency can be measured as a proportion of the best performance (least inputs per unit of output or most outputs per unit of input) observed in practice and devised a method for making these measurements given multiple inputs, outputs and constraints. This method is now usually referred to as data envelopment analysis (for excellent surveys of the literature on the use of data envelopment analysis and similar methods, e.g., substantively weighted least squares, to measure productive efficiency see Smith and Street, 2005; Mayston, 2003; Meier and Keiser, 1996). Farrell also proposed that identifying outstanding performers was the key to explaining performance.

Klitgaard and Hall (1973), frustrated by the inconclusive nature of the evidence on the determinants of school effectiveness, took up this proposition. They hypothesized that if, instead of looking at the average effect of institutional practices on student performance, they asked a different question, they might get somewhere. That question was: “Do some schools consistently produce outstanding students even after allowance is made for opportunity, perform most functions better than specialists (lean manufacturing), and that every step of the production process should be done perfectly (total quality management), thus reducing the need for rework (just-in-time delivery) and producing a higher quality end-product. Either way, the question arises, if manufacturers can see how Toyota does it, why can they not do it too? Neither lack of discipline nor the ego needs of bosses really seems like a satisfactory explanation.

The other alternatives are the best in theory and the best in history. Farrell’s work on the measurement of productive efficiency was something of a one hit wonder. He published 20 articles in JSTOR archived journals. Of those, his 1957/1962 articles received over 7,000 Google Scholar citations; none of the others got more than 60.
made for the different initial endowments of their students and for chance variation?”

Based upon analyses of data on Michigan, New York City, New York State, and Project Talent schools, they answered that question in the affirmative and then looked for differences in institutional practices that might account for the difference between unusually effective schools and typical schools. They found that the highly effective Michigan schools tended to have smaller class sizes, more well-paid teachers, and more experienced teachers. However, these relationships were not robust. Further analysis of their data failed to find similar relationships in New York City, New York State, or Project Talent schools. It also showed the identification of unusually effective schools was sensitive to the student achievement measures used and the period examined. Results for small schools were exceptionally unstable.

Representative Outputs. Good output measurement has four characteristics: the measures are comparable across different organizations, the measures are robust, the measures are easily replicable,\(^\text{16}\) and they should have a clear interpretation. Most economists have been content to use statistically adjusted test scores or value-added scores as outcome measures in their analyses of school and school district productivity, in large part because those measures are comparable across different organizations, inter-temporal instability can be overcome by extending the time period assessed (Kane and Staiger, 2002), and these measures, especially value-added scores, are easily replicated and interpreted. For example, in Figure 2 we use the Texas school district data set described below by Meier, Morton, and Molina (2010) to assess the performance of the

---

\(^{16}\)Ideally this means that the measures used should be designed and explained in such a way that they are numerically replicable for others without further help (provision of scores, own data sources and processing, own software, etc.).
~200 largest multi-school districts in Texas. Each district’s score equals the sum of its overall, White (non-Hispanic), Hispanic, and Black TAKS pass rates in 2009, less the sum of its overall, White (non-Hispanic), Hispanic, and Black TAKS pass rates in 1993. These scores have been arrayed in terms of overall TAKS pass rates in 2001, from left to right, lowest to highest. A straightforward best-practice analysis, would have us take the eight top districts or so in the sample and try to figure out how their practices differed from those of representative districts (Bowers 2010).

Figure 2: TAKS Based Multi-School District Effectiveness Scores

The main problem with what we have done here is that school districts, unlike automobile plants, produce a variety of outputs that are of interest, not only pass rates, but also a wide array of test scores, graduation rates, drop-out rates, college attendance rates, etc. The Texas school-district data set, for instance, includes 22 variables that can be used as output measures (we used four of them to construct Figure 2).

Three distinct analytical strategies can be used to address the multiple-output problem. None seems to be entirely satisfactory on all counts. The first is brute force: replicating one’s analysis using each of the output measures available to determine

---

17The array’s slope is probably due to top coding. As districts approach 100 percent on the TAKS pass rate their productivity improvement appears to stop although student test scores may continue to improve, college attendance rates increase or dropout rates fall.

18Obviously, this is a ubiquitous problem where statistical analysis is concerned and not only for best-practice analysis.
whether relationships between design features and outputs are sensitive to the choice of output measure. This is the most common approach taken by analysts of all stripes. It has the advantages of comparability, replicability, and ease of interpretation, at least when results are not, in fact, sensitive to the choice of output measures. Its drawback is that analysts must work through the full array of output measures in a conscientious and systematic manner for the results to be meaningful. This is time consuming and most don’t.

Output indices. Mayston and Jesson (1988) propose a second strategy for dealing with the multiple output problem: creating an index that comprehends a comprehensive battery of value-added and other relevant output measures. That is, in fact, precisely what most states have done to grade their schools and districts. However, state indices are largely ad hoc. Few if any have given very much thought to issues of normalization or component weights. Mayston and Jesson (1988) suggest that comparability requires normalization of component output measures to bring them to the same scale (preferably using representative samples). They further suggest that weights given to the components of an index should be calculated by (hierarchical) cluster analysis or by principal-component analysis. If principal-component analysis is used, they stipulate that weights should be corrected for the fact that some outputs are correlated and therefore contain the same information. In which case the recommended correction is to weight them inversely with the correlation structure.\footnote{This can be achieved by multiplying the data matrix by the inverse of the square root of the covariance matrix.} The drawback to this method is that if we want our indices to be comparable across districts, they probably won’t be easily replicated or explained. Or, if we want them to be easy to replicate and explained, they might not be
comparable (see, however, the performance indices used in Hanushek, 2010, and Mourghep, Chijioke and Barber, 2010).

*Output Surfaces.* Data envelopment analysis is a third strategy for dealing with multiple outputs. Ruggerio (1996) and also McCarty (1993) proposed a somewhat modified production function, which implies an efficiency frontier in the form of a n-dimensional ‘surface.’

\[ f(Y) \leq h(X|Z). \]  

(3)

Assuming convexity and monotonicity, data envelopment analysis can be used to construct this surface and to define a unique production possibility set for each district for a given environment in Z as:

\[ T(Z) \{ (X, Y) : f(Y) \leq h(X|Z) \}. \]  

(4)

This approach allows efficiency scores to be calculated for each district relative to its unique production possibility set. Using data from New York State Ruggerio (1996) calculated district efficiency scores and then regressed various design features on those scores to identify potential causal relationships. Unfortunately, 90 percent of the districts in his sample ended up on the efficient frontier (in large part because his production possibilities sets comprehended a wide array of inputs as well as outputs and constraints). Consequently, his regression analysis was inconclusive. Things that don’t vary cannot explain any variance. Banker, Janakiraman, and Natarajan (2003) used Texas school district data to construct a performance/cost frontier that included only (some) outputs and constraints and no inputs. Less than half of all Texas school districts were on the frontier and less than 1/3 of larger multi-school districts. Banker et al. (2003) could have
used their efficiency scores to investigate design issues, but chose not to, concentrating instead on cost-performance relationships.\textsuperscript{20}

In fact, for all the effort that has been devoted to frontier analysis, it is surprising how few memorable results this approach has produced. We attribute this outcome to its inherent drawbacks: its results are hard to explain and nearly impossible to replicate with any consistency. One can always obtain results – but those results are rarely stable.

\textit{The District Office as a Bundle of Institutional Arrangements: Insights from Political Science}

Relatively few political scientists have studied school district governance with an eye to student performance. Those, who have, have generally focused on the effects of institutional context. Terry Moe’s work is probably best known outside of political science.\textsuperscript{21}

Schools would perform better if they were smaller, more autonomous, and anchored in a strong sense of community. But schools did not get the way they are by accident. They are products of a complex system of political control, made up

\begin{footnotesize}
\textsuperscript{20}For example, collective bargaining for teachers in public school districts has no significant effect on student test scores, but materially increases per-pupil costs. In a second study, Banker, Chang, and Feroz (2008) use data from Minnesota school districts to compare parametric (OLS) and nonparametric (data envelopment analysis) methods. They conclude that severe violations of input linearity and separability assumptions render the use of parametric methods to explore cost and performance relationships impractical, let alone causal mechanisms. In contrast, they argue that data envelopment analysis is robust in the face of omitted variables and imperfect identification, where determinants of performance are ill specified and where input separability and linearity do not obtain. Unfortunately, imperfectly identified models are inherently open to interpretation.

\textsuperscript{21}Two of his articles have received over 1000 Google citations and 23 over 100.
\end{footnotesize}
of elected officials, administrators, and interest groups at all levels of government, in which the incentives are stacked in favor of large, bureaucratic schools. (Moe, 1994: 18; Chubb and Moe, 1990)

Moe’s research program has to do with the sources and consequences of interest-group power – mainly the influence of public employees on electoral outcomes and, thereby, public policies. Local-school districts provide a venue for his empirical research, permitting him to focus on specific kinds of employees and their behavior in elections they care about: teachers and school board elections.

Moe (2006) shows that teachers are between three and six times more likely to vote in school board elections than are non-teachers – although only when they live where they teach and have a personal stake in electoral outcomes – and are, as a consequence, quite successful at getting their candidates elected to local school boards. Moe (2009) further asserts that teacher influence over school boards is greatest where it is union organized and focused and, using California data, finds that, in large school districts, teacher control of school boards has a negative effect on academic achievement, particularly for minority students, perhaps because of teacher retention effects.

What can we conclude from Moe’s findings? We find the warrants for his claims plausible, even persuasive, but the evidentiary bases for his causal claims less so. The 2006 findings are based on a problematic sample and, in the 2009 article, it looks like he tortured the data until it confessed. Our main takeaway from Moe is that the electoral connection between teachers and school-board members results from the nonpartisan

---

22That doesn’t mean that he did or, even if he did, he’s wrong, merely that the descriptions of his data and methods show the earmarks of careful massaging. We await further analysis of these issues (Strunk and Grissom, 2010; Maranto and Wolf, 2010).
nature of school-board elections, which guarantees very low voter-participation levels. This tells us that it is very hard to predict the nature and direction of procedural political reforms. Otherwise, it is not at all clear what his findings imply, especially for teaching and learning.

The work of Ken Meier and Larry O’Toole (2011) is less well known outside of political science, but we conclude that it is much more important. Indeed, we predict that this volume will come to be seen as landmark contribution to the study of management. Good analysis combines good data, good theory, and credible identification. The analysis reported in this volume in a quantum jump forward on all three dimensions. It is almost entirely fortuitous that their subject matter is school district management.

First, Meier and his colleagues built a large-scale panel database (Meier, Morton and Molina, 2010) comprehending 17 years of data (1993-2009) for ~1,800 Texas school districts (including ~200 multi-school districts with ≥3000 students) featuring three components: district characteristics (83 variables, including 22 output variables), district resources (44 variables, including sources and uses of funds data), and district management, taken from a series of comprehensive surveys on styles, structures, processes, practices, goals and time allocations undertaken in 2000, 2002, 2005, 2007, and 2009 (see Appendix B), plus an additional survey concerned with how Texas district officials responded to the effects of Hurricanes Katrina and Rita, which shocked the region in 2005. The first two components are valuable, but not unique. Public access data on the variables included in them are available from a variety of sources for most states (although fewer years). The problem with most public-access district data is that is messy and requires a lot cleaning before it can be put to research uses. In contrast, the
management survey component of the Texas school district data set is absolutely unique. There is nothing quite like it in U.S. management literature. Combined with the first two components, the Texas district data set comprises a matchless analytical apparatus.

Second, O’Toole and Meier (2011) propose a new theory of government enterprise. To make predictions about behavior, we must presume that it is either habitual or purposeful. A model of organizational behavior based on the latter presumption necessarily starts with organizational purpose. For example, when we model business behavior, we usually start with the presumption that businesses maximize profit (or, more precisely, the present value of future free-cash flows thrown off by inventory turnover). Implicitly, Kiesling’s (1967) model (1) and its managerial variant (2) presume that school districts maximize productivity. Ruggerio (1996) recognizes that this isn’t a valid presumption and asks, instead, which districts look most like they are maximizing something (3, 4), given that we don’t know what it is? O’Toole and Meier (2011) take an altogether different tack. They presume that government enterprises maximize stability, subject, of course, to various constraints, including, perhaps, a productivity constraint. The underlying logic of their model is that an organization ought not to be a governmental enterprise, if we could afford for it to go out of business (Thompson, 2010).

Specifically, O’Toole and Meier (2011) posit an output vector at time $t$ or $Y_t$, and propose that current performance is a stochastic function of past performance at time $t-1$:

$$Y_t = s(Y_{t-1})$$

---

23 This presumption is usually justified by the observation that it is implied by the solution to any static optimization problem: revenue maximization, perquisite maximization, etc. This implication does not necessarily hold in the presence of uncertainty, over time, or under increasing returns.
In turn, function \( s \) reflects of a set of stabilizing features of organizational systems and buffers (standard operating procedures, budgets, civil service rules, and the like), \( O \),

\[
\text{Min } s_{var}\{Y_t = s(Y_{t-1}|O), s = h(O)\}
\]

Finally, because O’Toole and Meier’s theory is structured chronologically and because they are working with panel data, they can report not only correlation, but can also make strong claims about the direction of causation, that is: their claims are conclusive because they are properly identified.

Consequently, O’Toole and Meier (2011) provide compelling evidence for the significance of district organization and management. They report that the quality of district-level management accounts for up to 70 percent of the variance remaining, after controlling for nonmalleable factors in the district environment, in value-added student performance among Texas school districts.\(^{24}\) They further report that district performance is a straightforward, additive function of a set of malleable design features (e.g., superintendent quality and effort, district-staff size) and non-malleable variables (e.g., student diversity, the variety and complexity of funding sources), that different combinations of design features and variables can result in similar levels of performance, and that the key to district success lies in managing external relationships and allocating system resources, so as to buffer schools and teachers from external turbulence and shocks and stabilize instructional operations.\(^{25}\) O’Toole and Meier’s approach presumes

\(^{24}\) Note that between-district variance in average student value-added scores represents only 10 to 15 percent of the total variance in student test scores statewide.

\(^{25}\) Andersen and Mortensen (2010) take the O’Toole and Meier (2011) findings on the stability aspects of internal management and ask whether stability per se is an advantage for school districts. Their organizations are Danish schools, and they include a measure of resource stability in a model that predicts the performance of 140,000 school children on standardized tests. They find that stability in district resources contributes positively to
linearity and input separability or, to use the alternative language of organizational theory and design, it is non-configurative and equifinal (desired states can be achieved, from different starting points, via a variety of paths).  

One potential weakness of the O’Toole and Meier study is that it looks at only three aspects of the internal management of school districts and only indirectly: personnel and management stability (based on turnover ratios), the quality of human capital and its effective management (based upon the subjective assessments of district officials), and the ability to make good decisions, particularly when faced with major shocks (based on degree to which instructional programs were buffered from budget/enrollment shocks). They are silent on the issues that are most salient to education researchers, especially those concerned with the actual processes of teaching and learning and with using districts and inter-district partnerships and networks for large-scale improvement of instructional practice and student performance, although they acknowledge that best practice research in the field strongly emphasizes the need for dramatic change and development in these areas.

*The District Office as Strategic Apex and Technostructure: Insights from the Managerial Literature*

Even fewer students of business management have studied school district governance. In contrast to economists and political scientists, their work has generally

---

student performance over and above the level of resources or the change in resources. On the importance of stability to the implementation of educational reforms generally, see Desimone (2002).  

26 See also Meier et al. (2006).
focused on filling in the black box – on organizational culture, formal and informal relationships, alignment issues, and organizational learning. This work illustrates both the strengths and the weaknesses of management scholarship in general. Its strengths are thick description, an appreciation for the eventfulness of social processes, and openness to multiple conjectural causation. By multiple conjectural causation, we mean that the effect of organization and management is configuration (high performance depends upon the alignment of a cluster of design features – practices, attributes or activities, which are characterized by nonlinear and synergistic effects), unifinal (there is one optimal configuration, regardless of the presence or absence of nonmalleable factors in an organization’s environment), and that the most important design features (system improvement strategies, incentive structures, mechanisms for recruiting and evaluating personnel, professional development programs, and the allocation of decision rights to leaders, teams, and front-line workers) are internal, concerned primarily with the core technologies of the enterprise – in this case, teaching and learning. Its weaknesses are reliance on surveys and case studies, weak counterfactuals, and disregard of identification issues.27

---

27Pil and Leana (2009) and Leana and Pil (2006) is an exception to this generalization. They investigated the effects of teacher educational attainment and team horizontal tie strength and density on student performance in a sample of ~1,000 teachers organized into 239 grade teams. They found that both were significant predictors of performance. However, the effects of educational attainment depended on its relationship to the team purpose (core technology) and team strength. We think this finding should cause researchers to rethink the aggregation bias issue. Hanushek (1997) showed that studies that aggregate data into larger units of analysis (e.g., school- or district-level averages rather than using data on individual teachers) tend to show positive effects of teacher educational attainment. He concluded that these findings reflected ecological or aggregation bias. Pil and Leana’s findings suggest that it is also possible that we are looking at synergy. They conclude (2009: 1120) “For the more-able teachers, strong ties with peers are a key to unlocking these enhanced benefits both for themselves and for
Bill Ouchi, a leading expert on organizational culture and management control, is arguably the most prominent management scholar who has studied district governance.\footnote{His top four publications have altogether garnered ~8,000 Google cites; 18 of his publications more than 100.} Based upon a National Science Foundation funded comparative case study of six public school districts, which included 223 schools located in the six districts (the public schools of New York City, Los Angeles, Chicago, Edmonton, Seattle and Houston, the Catholic school districts of New York City, Los Angeles, and Chicago, and one independent school in each of the cities), Ouchi and Segal report (2003: 11): “If the district is run properly, all of the schools in it will be successful. If not, all schools will suffer, and only those principals who are willing to buck the central office will succeed.” Ouchi (2006) further claims that decentralization, delegating a high level of control over school budgets, staffing, schedule, and teaching methods, to principals is the key to improved student performance.

The best parts of Ouchi’s studies of district management are his thick descriptions of formal and informal relationships within schools and between personnel in district offices and principals, his normative reasoning about what constitutes a good outcome, and his historical narratives of events – where events are understood to comprehend both the actions taken by participants to develop or operate practices (the event’s trajectory) and resulting outcomes. His careful observations about culture, alignment issues, and organizational learning make compelling reading. What is least satisfactory about his
district-management work is the evidentiary basis for the conclusions he chooses to stress – his bullet points.

The McKinsey Study

The report on school-system governance (Mourshed et al. 2010) recently issued by McKinsy and Company, a management consultancy, is less compelling, but, on at least one count, conclusive and on several others, plausible. The authors of the report began by benchmarking. They identified systems that had “achieved significant, sustained, and widespread gains, as measured by national and international standards of assessment” (2010: 12). The criterion used was five years or more of consistent increases in student performance spanning multiple data points and subjects.

Then they did something unusual for best-practice analysis: they looked directly at program context. Instead of merely comparing a set of high performing systems to a set of representative systems, they focused on transitions from one performance level or development stage to another. To do so, they selected 20 systems from the set of high performing systems to obtain a diverse sample in terms of system sizes, locations, degree of centralization, sectors (public or private) and starting performance levels. They also devised a performance measure they could use to compare systems and their transitions from one level to the next. This involved normalizing 25 international and national assessments across multiple subjects and school levels, carried out between 1995 and 2010, to construct a single index that was keyed to the OECD’s PISA 2000 individual student test scores, with a mean of 500 and a standard deviation of 100. Using this index, they assessed the mean performance levels of the systems in their sample according to the

29Some, but not all, of these systems are districts.
following five categories: Excellent >560; Great 520-560; Good 480-520; Fair 440-480; Poor <440 at both the beginning and the end of the period studied. Next they identified and classified interventions aimed at performance improvement carried out by the systems in their sample. The 575 unique interventions they identified were sorted into 60 groups. Each group was categorized according to its area of impact (“for example, the ‘accountability area of impact’ includes the performance assessment, school inspection, and self-evaluation groups), kind of change (in resources, processes, or architecture), and its subject (e.g. principal, teacher, student). Finally, they performed a cluster analysis to determine which, if any, of the 60 intervention groups were ubiquitous and which, if any, were keyed to specific transitions. The result of that analysis is shown in Figure 3.

Figure 3: Interventions and Transitions
More generally, Mourshed, et al. (2010: 33-34) conclude that the educational development process is configurative and unifinal: “There is a common pattern in the interventions improving systems use to move from one performance level to the next, irrespective of geography, time, or culture.” However, different transitions call for different configurations or clusters of interventions. For example, there is an inverse correlation between the appropriate degree of centralization in the system and the transition level. Moreover, they claim that all effective transitions require interventions in six areas: “revising curriculum and standards, ensuring an appropriate reward and
remuneration structure for teachers and principals, building the technical skills of teachers and principals, assessing students, establishing data systems, and facilitating the improvement journey through the publication of policy documents and implementation of education laws.” However, they will manifest themselves differently in each of the transitions.

Finally, translated to American practice, Mourshed, et al. (2010: 92) argue that the proper functions of districts include: providing targeted support to schools/monitoring compliance; facilitating communication between schools and the center; encouraging inter-school collaboration; and buffering community resistance to change/securing stakeholder support

To sum up, we are convinced that level of performance matters: that making the transition from ‘poor to fair’ calls for different design features than going from ‘good to great.’ Indeed, there is a lot of food for thought throughout this report, but, absent satisfactory counterfactuals, it seems unlikely that, with respect to the efficacy of specific intervention configurations or role assignments, that the evidence presented in this report will persuade anyone not already inclined to be persuaded. Mourshed et al. (2010: 27) dismiss the absence of controls as irrelevant.

The systems that have been unsuccessful in trying to improve may carry out the same types of interventions that successful systems undertake – but there appears to be one critical difference, that they are not consistent, either in carrying out the critical mass of interventions appropriate to their performance stage, or in pursuing them with sufficient rigor and discipline.
This condition characterizes most configurative models.\textsuperscript{30} The problem results from a mismatch between the assumptions of these models and prevailing analytical methods. Fiss (2007) captures the methodological predicament succinctly: “the classic linear model treats variables as competing in explaining variation in outcomes rather than showing how variables combine to create outcomes. By focusing on the relative importance of rival variables, a correlational approach has difficulty treating cases as configurations and examining combinations of variables” (2007: 1181). Other methods work better, but econometric methods (interaction effects, simultaneous equation and hierarchical models, etc.) require fully identified specifications that posit the existence of well-behaved configurations of attributes while most other methods (clustering algorithms, factor analysis, etc.) produce inherently ambiguous results. Consequently, in most fields where configurative ideas are popular, evidence that performance depends upon adopting a specific pattern of practices, let alone aligning various design features with a particular context and/or strategy, is rarely if ever unequivocal.\textsuperscript{31}

Qualitative Comparative Analysis

Rizova (2011) proposes that there is an alternative analytical tool, Charles Ragin’s method (2008, 2000, 1994, 1987) of qualitative comparative analysis (QCA),\textsuperscript{32} which

\textsuperscript{30} And an even worse problem where causal relationships are configurative \textit{and} equifinal, as some scholars have suggested with respect to the effect of district organization and management on student learning (Honig, 2008, 2009).
\textsuperscript{31} A salient exception to this generalization is Brynjolfsson and Hitt (2000).
\textsuperscript{32} As a data analytical tool, QCA comes in different variations. The two most often used ones are denoted as $cs$QCA and $fs$QCA. The $cs$QCA technique (which is the originally developed QCA method) is appropriate for investigating causal sets of factors in “crisp sets” (i.e., when the attributes are categorized in dichotomous terms—they either exist or not; “high” or “low”; “success” or “failure”; “0” or “1”). Alternatively, when fine-
would allow researchers, particularly those concerned with school and district management, to explore how causal conditions combine to bring about outcomes of interest and, thereby, unravel the complexities of social and organizational realities. Once satisfactory hypotheses about configurative causal mechanisms have been identified, they could then be fruitfully tested on larger data sets using standard econometric methods.

QCA is grounded in the algebra of sets (binary data, “truth tables,” Boolean addition, multiplication and minimization, and combinatorial logic). Its underlying premise that systematic comparison across cases would allow researchers to discover the constellations of attributes related to outcomes of interest in small-N data sets (from at least 4 to 50) (Ragin, 2000) that are too large for traditional qualitative analysis and too small for conventional statistical analysis (Ragin, 1987, 2000, 2005, 2006; Rihoux and Ragin, 2009).  

The method involves the identification of two groups of cases: a benchmark group, comprised of the highest performers in the population, and a control group, randomly selected from the population remaining (less the worst performers in the population). The analysis proceeds by comparing configurations of cases to allow for the elimination of the non-causal conditions from the model. The empirical cases are then

\[ \text{QCA and csQCA software packages and manuals can be retrieved at http://www.fsqca.com.} \]

33In 1994, *Sociological Methods and Research* devoted a special issue to qualitative methods and included a number of articles concerning the application and advantages of QCA. More recently, scholars in the field of organizational strategy and management have successfully deployed the method to understand how design elements and practices influence organizational performance and outcomes (Fiss, 2007; Greckhamer et al., 2008; Herrmann and Cronqvist, 2009; Rihoux and Ragin, 2009; Rizova, 2007).
matched against the logical combinations. The result of this matching procedure is a list of plausible combinations of the independent variables associated with a particular outcome, in this case, “high” and “average” performance. Next, tests are run to evaluate the necessity and sufficiency of each causal combination and a Boolean minimization technique is applied to the ones that pass successfully. As a result of the minimization, the model is simplified and a final solution is produced – a combination of the variables associated with a particular outcome.

Rizova (2011) suggests that what we want to know is whether high-performing districts execute certain core projects differently from other districts and, if so, how? These are: designing system improvement strategies, designing and executing incentive structures for schools, principals, and teachers, recruiting and evaluating principals, providing professional development, allocating system resources, and buffering principals, teachers, and students from extraneous and distracting non-instructional issues. She concludes that in order to deploy QCA to answer this question we need: metrics for learning outcomes; a set of high performing districts and a set of representative districts; and an inventory of their design features and implementation processes. These are precisely the components seen in Mourshed et al. (2010) research, save only one – information on a set of representative systems; hence, although eminently feasible, Rizova’s proposal is yet to be carried out for school districts.

What’s New Here, if anything?

At the end of this review, we are left with the impression that economics, political science, and management literatures have not produced robust conclusions about the
relationship between district-level governance and student learning. Indeed, absent a
theory or understanding of how students learn, the challenge of explaining how
governance influences learning is challenging, to say the least. We know more about
learning today but we still do not know a lot. We explore systems of governance; learning
is a black box, and measure outcomes. The biochemical physiology and psychology, so to
speak, of learning is inside the black box. Hence, we are left with inferential reasoning,
however sophisticated the social science tools we bring to bear. The question though, is
how we can make our inferential reasoning more productive?

On this point, several of our conclusions are unexceptionable. First, we need to
build a national large-scale public-access district/school panel database on district
management practices. Aside from the Texas data set, nothing like this exists, so that the
returns to building such a data set would be very high. Second, we should support/run
more management field experiments: to uncover the causal impact of management
practices on district/school performance. The gold standard approach here is, in our view,
randomized control trials. Although these are expensive, it is difficult to see how the
evidence can be made secure without this type of approach. Third, given the extensive
experience of economists in large-scale data collection, natural experiments and field
experiments, they should be involved. Ongoing dialogue with other disciplines such as
organizational behavior, political science and sociology about measurement issues and
practices would also be helpful.

Our last conclusion is more debatable, but, we think, deserves a serious hearing.
Much of the best work in education administration shares the assumptions, strengths and
weaknesses of the work done by their counterparts in business management. Both
students of education administration and business management take many of their cues from John Dewey’s concept of experiential learning (Dewey, 1916, 1938) as explicated by various organizational theorists (Senge, 1990), especially Argyris and Schön (1996). From this perspective, organizational learning means collective problem solving, structured by a common set of expectations about what constitutes a good result that leads to the growth and development of common understandings about why things happen the way they do and, thereby, lead to large-scale improvement in organizations. This involves *communities of inquiry or practice* engaging in a cluster of behaviors, all of which require analysis and discussion of successes and failures, that add to the community’s pool of know how.

In turn, these assumptions reflect a shared view of the purpose of the research enterprise, one not necessarily shared with economists or other social scientists, who are often content with the acquisition of contemplative knowledge, that their job involves giving practical, useful advice about design and context features. We see this as a shared strength, in that it necessarily entails recognition that the subjects of their research are inherently complex and necessarily address disparate goals.

Their shared weakness is that, while their research is often informative, it is rarely conclusive. Murphy and Hallinger’s study is typical example of what we are talking about. Murphy and Hallinger (1986) looked at high-performing school districts in California, controlling for student composition, and found that their superintendents were especially knowledgeable about and the key initiators of changes in curriculum and teaching strategies; that their superintendents and system-level staff were unusually active in monitoring curriculum and instruction in classrooms and schools, as well as
active in the supervision, evaluation, and mentoring of principals; that their superintendents were also more likely to dismiss principals on the basis of their performance. They also claim that these districts showed a greater clarity of purpose, a greater willingness to exercise tighter controls over decisions about what would be taught and what would be monitored as evidence of performance, and a greater looseness and delegation to the school level of specific decisions about how to carry out an instructional program.

Similarly, Elmore (2004) concludes that high performing districts, especially high performing high poverty districts, are qualitatively different from others. Their leaders make their expectations for student learning both clear and urgent; they have challenging curricula and invest heavily in professional development; their teachers take collective responsibility for student learning; they analyze their practices and their results and, where current practices fail to achieve good results, they abandon them and try something else. These districts “are working hard to learn about their practice and beginning to focus on the individual and organizational conditions that create more powerful learning for adults and children.” He concludes that most districts “must learn how to improve teaching and learning in whole systems rather than isolated schools or classrooms.”

DuFour (2000) adds that “superintendents must help schools create the collaborative cultures that enable teachers to work together, engage in collective inquiry, and learn from each others' experience.” All of this makes sense. Indeed, these strictures have much in common with those of Ouchi (2006) or Mourshed et al. (2010). But they are also no more conclusive. Students of educational governance, like their counterparts in business management, really need to think more about causal mechanisms, building
and testing properly identified models, and not rest content with reverse engineering implementation processes.

Finally, there is the matter of making the kinds of problem addressed and the would-be solutions to them clear, both in the conduct of research and in its explication. When economics or one of the other social sciences is the reference discipline, researchers tend to focus on stocks and flows of assets and resources, liabilities and disadvantages that favor or inhibit program realization. When management is the reference discipline, researchers tend to be concerned with finding knowledge that can be used to overcome certain stylized challenges, such as controlling production, managing the delivery process, developing strategy, or creating capabilities for future delivery. Some might see this merely a matter of presentation. We would not entirely dismiss that interpretation, especially where social scientists formulate theories about carrying out functions that promote programmatic success. Nevertheless, presentation matters both for replication and for extrapolation. One should make the purpose of one’s enterprise clear from the outset.
Appendix A

Our survey relied on Google Scholar. We first searched on the following keywords: school district, board of education, school superintendent, and school-district office, with the search restricted to the top twenty general and education relevant special-purpose journals in economics, political science, and management. As a result we identified 229 articles, sorted according to relevance to our subject.

### Economics

<table>
<thead>
<tr>
<th>Journal of Economic Review</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brookings Papers on Economic Activity</td>
<td>1</td>
</tr>
<tr>
<td>Econometrica</td>
<td>3</td>
</tr>
<tr>
<td>Journal of Econometrics</td>
<td>5</td>
</tr>
<tr>
<td>Journal of Economic Literature</td>
<td>4</td>
</tr>
<tr>
<td>Journal of Economic Perspectives</td>
<td>12</td>
</tr>
<tr>
<td>Journal of Labor Economics</td>
<td>5</td>
</tr>
<tr>
<td>Journal of Political Economy</td>
<td>4</td>
</tr>
<tr>
<td>RAND Journal of Economics</td>
<td>0</td>
</tr>
<tr>
<td>Review of Economic Studies</td>
<td>1</td>
</tr>
<tr>
<td>The Quarterly Journal of Economics</td>
<td>7</td>
</tr>
<tr>
<td>The Review of Economics and Statistics</td>
<td>9</td>
</tr>
</tbody>
</table>

| Journal of Public Economics         | 12|
| Journal of Human Resources          | 13|
| Journal of Law, Economics and Organization | 1 |
| Journal of Law & Economics          | 1 |
| International Tax and Public Finance | 3 |
| Journal of Urban Economics           | 9 |
| Economic Inquiry                    | 4 |
| Economics of Education Review       | 44|

Total: 147

### Management

| Academy of Management Journal       | 6 |
| Academy of Management Review       | 1 |
| Administrative Science Quarterly    | 7 |
| Advanced Strategic Management      | 0 |
| California Management Review       | 0 |
| Decision Sciences                  | 0 |
| Group and Organization Management  | 6 |
| Harvard Business Review            | 2 |
| International Journal of Management Reviews | 0 |
| Journal of Information Technology  | 2 |
| Journal of Management              | 4 |

Total: 147

43
| Journal of Operations Management | 1 |
| Management Science | 2 |
| MIS Quarterly | 0 |
| Omega | 1 |
| Organization Science | 1 |
| Organization Studies | 1 |
| Organizational Behavior and Human Decision Processes | 2 |
| Organizational Research Methods | 0 |
| Strategic Management Journal | 1 |

<table>
<thead>
<tr>
<th>Political Science</th>
<th># Articles about district governance and/or educational performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Journal of Political Science</td>
<td>3</td>
</tr>
<tr>
<td>American Political Science Review</td>
<td>0</td>
</tr>
<tr>
<td>Journal of Policy Analysis and Management</td>
<td>9</td>
</tr>
<tr>
<td>Journal of Politics</td>
<td>2</td>
</tr>
<tr>
<td>Journal of Public Administration Research &amp; Theory</td>
<td>10</td>
</tr>
<tr>
<td>Journal of Theoretical Politics</td>
<td>0</td>
</tr>
<tr>
<td>Law &amp; Society Review</td>
<td>0</td>
</tr>
<tr>
<td>Legislative Studies Quarterly</td>
<td>0</td>
</tr>
<tr>
<td>Political Analysis</td>
<td>0</td>
</tr>
<tr>
<td>Political Behavior</td>
<td>1</td>
</tr>
<tr>
<td>Political Research Quarterly</td>
<td>2</td>
</tr>
<tr>
<td>Political Science Quarterly</td>
<td>2</td>
</tr>
<tr>
<td>Politics and Society</td>
<td>1</td>
</tr>
<tr>
<td>PS: Political Science and Politics</td>
<td>0</td>
</tr>
<tr>
<td>Public Administration Review</td>
<td>12</td>
</tr>
<tr>
<td>Public Opinion Quarterly</td>
<td>0</td>
</tr>
<tr>
<td>Social Forces</td>
<td>1</td>
</tr>
<tr>
<td>Theory and Society</td>
<td>0</td>
</tr>
<tr>
<td>Urban Affairs Review</td>
<td>1</td>
</tr>
<tr>
<td>Urban Studies</td>
<td>1</td>
</tr>
</tbody>
</table>

These are arguably the top twenty journals in each field that might conceivably publish articles bearing on our topic: the influence of district governance on teaching and learning. This means that we omitted journals specializing in topics such as finance (economics), international relations (political science), and logistics (management), for examples. In economics this procedure led us to omit a significant number of journals. Consequently, we moved down the list and added the top-ranked journals that published articles concerned with education policy and practices – although, as it turned out, not necessarily articles related to our topic.
The rankings shown above primarily reflect ISI impact and article influence scores. Citation-based studies are the norm in economics and management. Within political science the dominant approach to journal assessment has been reputational. Consequently, we checked our list against Garand & Giles (2003). About the same journals appear, although not necessarily in the identical order.

This exercise demonstrated three things. First, there are a lot more high-impact journals in economics than in political science or even management. Second, top journals in economics publish much more about education then do top journals in political science or management. For example, the American Economic Review, the American Political Science Review, and the Academy of Management Review, are arguably the top journals in their fields. According to Google Scholar, the American Economic Review published 24 articles mentioning the terms ‘school boards’ and ‘learning,’ the American Political Science Review published 9, and the Academy of Management Review 3. Moreover, nine of the articles published in the American Economic Review are concerned with the determinants of educational performance, although those determinants rarely include district level policies and practices. Only one article in the flagship journals of the other two fields is concerned with either district governance or educational performance, Sutton & D’Aunno (1989). This article looks at the effects of the loss of financial resources and work force reduction on a sample of school districts. Sutton & D’Aunno show that they cause mechanistic shifts in organizational structures and jobs. They note that these kinds of changes often adversely affect organizational performance, but did not investigate their effects on student learning in the districts they studied. Third, when economists look at education, they are somewhat more likely to take learning outcomes as their dependent variable than are political scientists or management scholars, although as a general rule they all do so rather uncritically.

Next, we pursued the sources they cited to identify the article or articles that initiated a significant research path (the Ur-text or texts). Finally, we followed each of these paths back to the present. Our survey describes the start of each path, identifies any subsequent findings confounding initial results, and then skips ahead to recent work at the cutting edge of the research enterprise. Finally, in most cases, we chose to review a more comprehensive text where one was available dealing with the material reported in one or more articles published in our primary list of journals (e.g., Ouchi’s book rather his articles in Academy of Management Journal and Organizational Science and Klitgaard’s Rand report or Moursched’s McKinsey report rather than their much shorter articles appearing in Quarterly Journal of Economics and Harvard Business Review).
Appendix B

Figure S.1: Steps for Policymakers in Examining the State’s Educational Governance System

<table>
<thead>
<tr>
<th>WHAT are the goals of the system in terms of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Structure and organization</td>
</tr>
<tr>
<td>• Finance and Business Services</td>
</tr>
<tr>
<td>• Human Resources/Personnel</td>
</tr>
<tr>
<td>• Educational Programs</td>
</tr>
</tbody>
</table>

| WHO is best situated to carry out the tasks necessary to meet those goals? Think about institutions and individuals at the various levels of the system (e.g., Governor, Legislature, SBE, SPI, CDE, District Superintendents, District Boards, County Offices of Education, Principals and Teachers) |

<table>
<thead>
<tr>
<th>HOW should these institutions or individuals best induce others to implement policy? What mix of the following is best suited to meet the goals:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Mandates</td>
</tr>
<tr>
<td>• Inducements</td>
</tr>
<tr>
<td>• Capacity-Building</td>
</tr>
<tr>
<td>• System-Changing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evaluate. How does the system rate in terms of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Stability</td>
</tr>
<tr>
<td>• Accountability</td>
</tr>
<tr>
<td>• Innovation, flexibility, and responsiveness</td>
</tr>
<tr>
<td>• Transparency</td>
</tr>
<tr>
<td>• Simplicity and efficiency</td>
</tr>
</tbody>
</table>

Source: Brewer and Smith 2006: ix
Appendix C

Description of the Texas SD surveys

In addition to data on superintendents’ characteristics, such as, highest degree earned, tenure on the job, number of years with the district, age and previous jobs, the survey asked five sets of questions about the management of school districts.

The first set of twenty-three questions asks about the behavior of district leaders with regard to: the frequency of their interactions with key stakeholders (school board members, local business leaders, parents, state legislature, teachers’ association, etc.), the time they spend dealing with thirteen separate core managerial issues related to: personnel, curriculum and instruction, financial management, legal matters, strategic planning, individual campus and student issues, and such.

The second set of questions asks about superintendents’ roles, practices, and preferences with respect to matters such as adopting new ideas and practices, conflict over matters of education within the community, degree of influence of the state legislature on the work of the district, and whether or not the superintendent should be involved in curriculum selection and in shaping school district policies. This section also asks the leaders to identify the six most important problems facing their district.

The third set of questions asks superintendents to assess the quality of their district’s facilities, teachers, professional development, support from the community, as well as of the involvement of parents and the school board.

The penultimate set of questions asks about ethnic and cultural diversity – availability, and quality of programs for disadvantaged students, promotion and retention of people of color, frequency of diversity training, etc.

The final set of five questions explores the views of the superintendents on the stability, complexity and uncertainty of the district’s political, social and economic environment and its effects on operations and outcomes.
References

Andersen, F. Ø. (2010). Danish and Finnish PISA Results in a Comparative, Qualitative Perspective: How Can the Stable and Distinct Differences between the Danish and Finnish PISA Results Be Explained? *Educational Assessment, Evaluation and Accountability*, 22(2), 159-175.


Amsterdam: Elsevier.


Brown, J., & Duguid, P. (1991). Organizational Learning and Communities of Practice:


Downloaded November 12, 2010.


the Operational Research Society, 54(7), 679-691.


