The Practical Relevance of Accountability Systems for School Improvement: A Descriptive Analysis of California Schools

CSE Report 713

Heinrich Mintrop

National Center for Research on Evaluation, Standards, and Student Testing (CRESST) / University of California, Berkeley

Tina Trujillo

National Center for Research on Evaluation, Standards, and Student Testing (CRESST) / University of California, Los Angeles

April 2007

Center for the Study of Evaluation
National Center for Research on Evaluation, Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 1.1 Comparative Analyses of Current Assessment and Accountability Systems; Strand 4: The California School Accountability System and the Improvement of Low-Performing Schools
Heinrich Mintrop, Project Director

The work reported herein was supported under the under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education.

THE PRACTICAL RELEVANCE OF ACCOUNTABILITY SYSTEMS FOR SCHOOL IMPROVEMENT: A DESCRIPTIVE ANALYSIS OF CALIFORNIA SCHOOLS

Heinrich Mintrop
CRESST/University of California, Berkeley

Tina Trujillo
CRESST/University of California, Los Angeles

Abstract

In search for the practical relevance of accountability systems for school improvement, we ask whether practitioners traveling between the worlds of system-designated high- and low-performing schools would detect tangible differences by observing concrete behaviors, looking at student work, or inquiring about teacher, administrator, or student perceptions. Would they see real differences in educational quality? Would they find schools that are truly more effective? In this study, we compare nine exceptionally high and low performing urban middle schools within the California accountability system. Traversing the nine schools, our travelers would learn that schools that grew on the state performance indicator tended to generate internal commitment for the accountability system. They eschewed the coercive aspects of accountability, maintained a climate of open communication, and considered the system as an impetus for raising expectations and work standards. On the instructional side, this commitment translated into the forceful implementation of structured language arts and literacy programs that were aligned with the accountability system. If our travelers expected to encounter visible signs of an overall higher quality of students' educational experience in the high-performing schools, they would be disappointed. Rather they would have to settle on a much narrower definition of quality that homes in on attitudes and behaviors that are quite proximate to the effective acquisition of standards-aligned and test-relevant knowledge.

This paper grows out of two motifs that have repeatedly surfaced in conversations with school practitioners and students in the Principal Leadership Institute (PLI) at the University of California, a program in which one of the authors has been an instructor for many years. The two motifs speak to the practical relevance of school accountability systems for school improvement.

Five years ago, when asked to introduce their school, PLI students would name their school's demographics, the likeable features of school life, perhaps the relationship between principal and teachers, and end up describing some major challenges. In 2005, after 6 years of state accountability, PLI students most often state that they are a 1-1 school; a 3-5 school; a school that grew 50 points last year. What they are referring to is the Academic Performance Index (API), the state's prime indicator for school quality, and the state and similar schools ranks that are computed based on this indicator. Within 5 short years those numbers have

seemingly become signals of a school's quality and character, an increasingly powerful shorthand and social fact in the lives of school practitioners. Is this justified?

Increasingly, administrators and teachers in 1-1 schools are urged to avail themselves of lessons to be gleaned from the practices of 3-5 or 1-10 schools that have presumably mastered similar educational challenges with higher success. But we frequently encounter the assertion from practitioners in schools classified as lower performing that "we have looked at these schools, but we already do all of the things they do; and they don't look that different from us." Is this self-serving?

The two motifs are paradoxical: Either the system-based performance categories stand for some broader characteristics of school quality, or they are disconnected from educational and organizational conditions sufficiently tangible for educators to learn from. Our research explores this paradox. It takes from the conversations with practitioners that a connection cannot be taken for granted and yet such a connection is intuitively made. Naturally, when performance indicators and practitioner experience of quality and effectiveness speak to the same reality, practical relevance of accountability systems is heightened. In reporting our findings, we adopt the lens of some imagined practitioners who, based on information from the state accountability system, travel to a number of schools in order to learn what to do. We accompany these travelers in our role as researchers assisting with robust instruments and the benefit of systematic inquiry.

In this paper, we first describe the state context of the study; then define in more detail what we mean by practical relevance; lay out the models and measures that guide our assumptions about educational quality and organizational effectiveness; explain the design of the study; document the development of robust instruments; and finally present our findings. In writing this paper, we have three purposes in mind. We want to facilitate practitioner discussions with evidence and rational argument, draw lessons for good system designs, and convince the research and policymaking public of the necessity for more of the kind of validation research that we are conducting here.

The California Accountability System

The model of organizational development that underlies the current accountability system in California and the federal No Child Left Behind (NCLB; 2002) legislation revolves around state standards, assessments, and goal setting

based on a small set of quantitative indicators, most notably test scores condensed in the Academic Performance Index (API) and, since the passage of NCLB, Adequate Yearly Progress (AYP). Easily readable numbers presumably make school performance transparent and steer schools' efforts towards the attainment of continuous growth. Looming sanctions fill in where the softer touch of steering by quantitative goals has presumably not taken hold with sufficient force.

In 1999, California followed in the footsteps of other states with the installation of its own outcome-based accountability system (Mintrop & Trujillo, 2005). The state calculated a performance score from year to year based on a formula of weighted tests and computed yearly target API scores that were to gradually move all schools to the level of 800 API points (California Department of Education [CDE], 2006). Because growth targets were calculated as 5% of the difference between a school's API and the statewide target of 800, lower performing schools were expected to meet larger growth targets than their higher performing counterparts. A set of rewards, sanctions, and supports were to reinforce accountability demands.

The validity of the California accountability system has been discussed on a number of grounds:

- Since 1999, the system's consistency and the robustness of the API have been questioned as types of tests used, and the weights attached to them have changed dramatically, away from norm-referenced tests, such as the Stanford 9, to the California Standards Tests (CST) that are criterion-referenced, subject-matter based, and better aligned with state standards (Russell, 2002).

- School improvement trajectories are plotted through average student performance from year to year, falsely assuming population stability over time (Linn, Baker, & Betebenner, 2002)—a particularly acute challenge for middle and junior high schools that turn over their entire population every 2 or 3 years.

- Many English language learners, abundant in the state's schools, are tested in a language medium that they poorly understand (Abedi, 2004).

- Measurement errors make classifications for schools with typically small API movements uncertain (Kane & Staiger, 2002).

- The likelihood of failure increases with the number of statistically significant subgroups (Linn, 2005).

- AYP as a second way of calculating school performance has clouded the presumed clarity of incentives and goal setting based on the state's own API (Kim & Sunderman, 2005).[1]

- Gaming, grade retention, and push-out of weaker students may occur (Haney, 2000).

- Strong correlations between school and district failure and the socioeconomic status of the community they serve point to societal factors beyond the control of school leaders. Output expectations are not calibrated to input requirements (Oakes, 2002).

While these problems may put the validity of the California system and its central indicator, API, in question, we begin our study with the assumption that every system of measuring complex performance contains errors and distortions, but these may be tolerable if gains and losses measured by narrow system indicators match up with something real, more broadly-based, and concrete that has practical relevance and that can be exploited by reformers to good ends. For example, the shift from norm-referenced tests to the state standards tests, while potentially increasing measurement error, may have actually enhanced the system's practical relevance as assessments are now more closely aligned to taught curricula, enabling publishing companies and support providers to produce curricular materials and provide consulting services better connected to tested school performance.

The Practical Relevance of Accountability Systems for School Improvement

Accountability systems sort schools according to winners and losers. Losers are schools that perform at below-average rank, stack up negatively against schools with similar demographics, or have repeatedly not met their targets. Losers are classified as *underperforming*, *program improvement*, or *corrective action* schools and face pressures and sanctions. Winners manage to exceed their growth targets in absolute terms or compared to demographically similar schools. These high-growth schools are touted as role models (Carter, 2001; Haycock, 1999; Reeves, 2000) and are eagerly sought out as models of improvement by some less fortunate, low-growth or declining schools. As the accountability system becomes more and more institutionalized, performance indicators, such as the California API, have come to

[1]At the time of data collection for this study, AYP was beginning to make inroads into the practical life of schools, though the API was still the more established measure and preferred orientation. It encapsulated the state's authoritative judgment of school quality and improvement success and was deemed a better fit with school reality than the ominous AYP with its presumably hard-to-reach proficiency targets and accelerated sanctions.

confer public value on schools and have entered educators' minds, if not hearts. As reaching target scores is of paramount importance to schools' organizational survival and standing in their districts and communities, the system is presumably attaining increasing evaluative and self-evaluative power.

The justification of this power, as far as it concerns the business of school improvement, seems to hinge on two fundamental claims:

- The state performance scores are valid inferences of school quality; and
- High growth in these indicators over time is not a chance occurrence, but a reflection of superior school improvement efforts on the part of high-performing schools that can be emulated by lower performing counterparts.

These two claims not only speak to the validity of accountability systems in a more narrow sense, but are of utmost practical relevance, particularly in and around schools that traditionally occupy the bottom rungs of the social and educational status hierarchy and find themselves in dire need of ideas on how to improve.

Capitalizing on the presumed practical relevance of the system, educational reform organizations in California and elsewhere (EdSource, 2003; Springboard Schools, 2005; WestEd, 2005) have studied schools with high API or API growth. Following a classic outlier design, these studies tend to showcase such schools' presumed exemplary practices. While these kinds of studies are useful, they are also limited. Most important, they presuppose the practical relevance or validity of system performance indicators, something that begs for substantiation.

We use the term validity in this paper loosely. In asking what social consequences a measure of key importance, such as API, has for its central design purpose (i.e., school improvement), we have a notion of "consequential validity" (Messick, 1988) in mind. But as nonspecialists in the field of measurement, we prefer the term "practical relevance." As school improvement researchers, we deliberately take on the accountability system as received by lay practitioners, not as intended by statisticians or technical system designers. In our search for practical relevance, we ask whether practitioners traveling between the worlds of API-designated high- and low-performing schools would detect tangible differences by observing concrete behaviors, looking at student work, or inquiring about teacher, administrator, or student perceptions. Would they see real differences in school quality? Would they find schools that are truly more effective?

Our definition of practical relevance (or consequential validity, if you will) hinges on five conditions. In outcome-based and indicator-driven accountability systems of high practical relevance, indicated high-performing schools are distinguished from their lower performing counterparts by (a) having higher achievement as measured by the performance indicator(s); (b) facing similar educational challenges; (c) providing a higher quality educational experience for students; (d) functioning more effectively; and (e) engaging with accountability mechanisms in a way that brings about aforementioned desirable properties. Oftentimes, we assume that high-performing schools, indicated as such by their accountability system, satisfy all five conditions. At times, when our initially mentioned practitioners speak of their schools with shame or pride as 1-1 versus 3-3, or 643 versus 517 schools, they seem to say that the system-indicated rank connotes this broader spectrum of conditions.

A study that is designed along these five conditions must do the following: identify schools or groups of schools with sufficiently distinct performance scores and close similarity with regard to their demographic background; compare schools classified as "high" with ones classified as "low" so as to discern what is truly unique about the former; and connect measured achievement to the quality of students' educational experience, organizational effectiveness, and engagement with the accountability system.

Once proper samples are drawn or cases selected, the exploration of schools is guided by three main research questions:

1. Is absolute performance level, as measured by the state indicator, matched by other quality criteria, detectable in the concrete life of a school, that speak to a broader view of school quality?
2. Is growth over time, as measured by the state indicator, associated with the presence, absence, or strength of school characteristics that have shown to play a prominent role in effective school improvement?
3. Are the specific mechanisms of the accountability system instrumental for school improvement?

Models of Quality and Effectiveness

How educators define and imagine the good school has consequences for the way they keep school and teach, but also for the dynamics of organizational development they engage in. A notion of quality that centers on the big ideas of human knowledge, cultural relevance, teachers as founts of knowledge, intense personal relationships between learners and teachers, or community accountability

will set different accents in organizational development than one that centers on proficiency, curriculum alignment, teachers as transmitters of knowledge, or test-driven accountability. But the relationship between notions of educational quality and dynamics of organizational development is bi-directional. While in earlier historical periods commitments to notions of educational "goodness" may have driven organizational change dynamics, as in the "school wars" of the sixties (Ravitch, 1983), the current phase of high-stakes accountability seems to rest on an inversion of this relationship. It does not seem to be fueled as much by ideological or moral zeal about aims as by certitudes about principles of organizational development and productivity that are borrowed from the world of business. Thus, ideas about organizational development drive the educational aims and notions of quality that come into prominent view.

There has been considerable debate, however, about this relationship. The Texas system has received the most attention in this regard. (For examples of this debate, see the collection of articles in Skrla and Scheurich, 2003.) There are those who find high performance, as indicated by the system, associated with raised expectations, streamlined operations, and a new urgency and determination among educators and system administrators; that is, they make their argument on the grounds of achievement and organizational effectiveness. Others argue that schools tighten up in the service of the wrong aims: unduly concentrating on test preparation, narrowing the curriculum to the most basic of skills, abandoning deep understanding in favor of superficial coverage, becoming overregulated and undemocratic places, and excluding statistically disadvantageous children. For these critics, the quality of the educational experience has deteriorated and the accountability system has ultimately produced unwanted organizational dynamics.

Our study touches on this important debate by employing measures of quality that pick up on the concerns of critics and by homing in on the connection among measures of educational quality, organizational effectiveness, attitudes to accountability, and implemented change strategies.

Educational Quality

The utility of accountability systems for school improvement increases when the system's success cases (in this case, relatively higher API) also rate highly on other important quality indicators not measured by the system. We gain a practical understanding of school quality that moves beyond standardized testing if we know

how students experience their school, what kind of teaching they encounter, and what kind of work they produce in their classrooms.

In our research, we capture students' educational experience with the help of three sets of data: students' perceptions of their school, observations of lessons and enacted curricula, and student writing samples. We ask about students' perceptions of academic engagement, academic press, teacher care, peer collaboration and safety. These are conventional variables that have shown to be of particular relevance in previous school effectiveness and large-scale student achievement studies (e.g., Newmann, Bryk, & Nagaoka, 2001; Organisation for Economic Co-Operation and Development, 2000; Teddlie & Reynolds, 2000). "High-quality" schools are defined as ones in which students feel engaged and challenged, but at the same time safe and cared for. Collaboration among students is an important civic component of good schools as well.

For lesson observations and student work samples we concentrate on English language arts because all of the schools we studied put the overwhelming focus of their improvement efforts on this subject and more generally on literacy development. We hypothesized, again in line with a long tradition of effective teaching research (Scheerens, 1992), that high quality teaching is characterized by effective use of time, empathic and active teaching, and a variety of cognitively complex learning activities. In evaluating the quality of student writing samples, we explored basic writing skills (e.g., clarity, coherence, language conventions), but also wanted to ascertain degree of complexity in constructing arguments and interpreting phenomena (Newmann, Smith, Allensworth, & Bryk, 2001; Newmann & Wehlage, 1995). Lastly, we use student suspension rates in conjunction with perceptions of safety as a proxy for student discipline. Thus, with our quality indicators we explore how schools are doing with regards to basic order and basic skills, but also with regards to more advanced and complex learning and attitudes.

Organizational Effectiveness

Following the research on school effectiveness and school improvement, we hypothesized that if the accountability system worked properly we would find organizational characteristics in the system's success cases (i.e., schools with high growth in API over time) that are commonly believed to hold sway in effective schools. Indeed, accountability systems are designed to foster the development of school change along the lines of the effective schools model. Research on effective

schools and school improvement has identified a large variety of factors (Sammons, 1999; Scheerens & Bosker, 1997) out of which we selected a few salient but fairly conventional ones that appeared most relevant to extant conditions of schools under accountability. The ground was prepared for this work during an earlier study on "schools on probation" in the states of Maryland and Kentucky directed by one of the authors (Mintrop, 2003, 2004).

Time and again, the research has pointed to the centrality of leadership for school improvement success. In the literature on effective schools and improvement, the principal appears in several guises: as capable *manager*,[2] skillful *instructional leader*, or credible *moral leader* (Deal & Peterson, 1991; Fullan, 2003, 2005; Hallinger & Heck, 1996; Sergiovanni, 1992). In some conceptions of school change, the principal appears as *supportive*, fostering a climate of respect for professional *autonomy* and *open communication*; other conceptions emphasize his or her role as strong initiator and enforcer of rules, particularly in often chaotic urban high-poverty environments (Teddlie & Reynolds, 2000). While there seems to be a tendency for some leaders, coming under strong accountability pressure, to increase *control* and reinforce the system's *urgency* and pressure (Mintrop, 2004), others may create momentum for collective problem solving. Many strong leaders seem to combine managerial, instructional, and moral aspects into their role.

Key characteristics of faculty culture can be captured in the tension between unity and flexibility. Cohesive *collegiality* around common sentiments or purposes and a *learning orientation* that maintains continuous openness may in some faculties go hand in hand, and in others conflict with each other (Achinstein, 2002; Little, 1982; McLaughlin & Talbert, 2001). Under conditions of accountability, *pulling together* and assuming responsibility by adhering to *norms of performance* (Elmore, 2004) seem to be especially salient characteristics for faculties that are in need of a collective response.

Effective schools require a motivated work force. No change process can get off the ground and be sustained without teachers' high *involvement*, and it is no secret that in many urban schools, classified as low performing, meeting external expectations of continuous growth requires a high energy level and time commitment from teachers (Ingersoll, 2001; LeCompte & Dworkin, 1991), one that may go beyond contractual hours. *Hard work* is required. But challenge (and

[2]Words in *italics* directly refer to variables measured or explored.

concomitant stress) needs to be balanced with a sense of *satisfaction* with one's work and the expectation that one can succeed. Otherwise, *morale* may be low and *commitment to stay* in the challenged environment may be reduced (Odden & Kelley, 1997; Rowan, Chiang, & Miller, 1997). Expectation of success may in good measure be dependent on one's sense of *instructional efficacy* (Ashton & Webb, 1986; Hannaway & Chaplin, 1994), for example, in managing the classroom or reaching difficult children. Under conditions of group accountability, a sense of *colleagues' skills* and *test-related efficacy*, requisite for success of the school in the system, may play a role as well.

Recent literature on school improvement has pointed to the salience of a school's instructional program and management for a school's chance to improve. We selected *instructional program coherence* (Newmann, Smith, et al., 2001), *data use* for evidence-based decision making (Blankstein, 2004), school improvement *planning* (Mintrop & MacLellan, 2002), and a *strategic orientation* towards change (McBeath & Mortimore, 2001) as aspects that capture the presumed rationality of the accountability-driven change model. Based on the increased prominence of the central office in school improvement efforts (Hightower, Knapp, Marsh, & McLaughlin, 2002), we also included the *district instructional and operational system* as an external source of change.

Responses to Accountability

The practical relevance of an accountability system is naturally tied to how schools pick up on the system's signals and messages. Schools could maintain a posture of defensiveness against unwarranted external demands or may develop an orientation of constructive engagement (see Mintrop, 2004, for a more extensive discussion). Accountability systems are of high practical relevance if pressures, incentives, directives, and flows of information emanating from, or generated by, the accountability system have played a key role in the life of high-performing schools. Alternatively, schools may have paid no attention to, or have improved in opposition to, the system. Change may have occurred naturally (Teddlie & Stringfield, 1993) without the influence of external levers; or high-growth and low-growth schools may have paid similar attention and made similar use of the potentially motivating force of accountability mechanisms without achieving the same results.

We base our exploration of schools' responses to accountability on the following model (Mintrop, 2004). Schools attach varying degrees of *goal importance* to the demands of the accountability system. Importance could be more externally or internally motivated (Deci & Ryan, 1985). In an external nexus, teachers could calculate extrinsic rewards, such as enhancement of professional prestige or aversion of disadvantages; that is, they would act primarily out of a sense of *external validation*. They may also accept the state government's normative authority, or *authoritativeness*, to give teachers directions, in specified areas or more generally. Less benign than the appeal to one's sense of loyalty or desirability of reward is the experience of coercive power. Accountability systems can create *pressure* and an imminent sense of personal sanctioning and *threat*.

In contrast to these primarily external motives to heed accountability demands there could be more internalized motives. Usefulness of the system in providing *focus* within the uncertain technical culture of teaching and the traditional legitimacy of testing as enhancing *diagnostic* capacity inhabit the outer layers of internalization. *Validity* and *fairness* connote a deeper sense of rightful judgment. Usefulness, rightfulness and *realism* of targeted goals are the tripod on which the effectiveness and steering capacity of a performance indicator rest (Fitz-Gibbon & Kochan, 2000). They are the prime sources of meaningfulness. If accountability systems worked properly, teachers would supposedly have *raised expectations* for their students' performance and the caliber of their own work. If they internalized the system properly, they would experience stronger *goal integrity*, that is, a better match between system demands, needs of students, and their own values.

Various combinations between leadership, faculty culture, and response to accountability are possible. Two extreme scenarios shall be described for heuristic purposes. The primary mechanism of accountability power in a given school may be pressure and threat. Principals may seize upon these pressures, reinforce urgency or even fear, increase control, and tighten up the organization and instructional program. By contrast, schools may accept the accountability system as a meaningful guide; principal leadership may foster a culture of organizational learning among staff members (Louis & Kruse, 1998) that in turn reinforces commitments to common instructional goals and structures. The reality will likely be mixed (Louis, Febey, & Schroeder, 2005). For this study, the power of accountability systems could rest on external validation, authoritativeness, coercion, usefulness, rightfulness, or a

mix thereof. Human beings, as we know, respond to all of these motives under given circumstances. (Core variables for this report are listed in Appendix A.)

Methods and Data

There are several ways to explore the practical relevance of school accountability systems. Our way is to look at a relatively small number of schools in depth that differ in absolute performance and growth over time as measured by the prime state indicator, API. We studied nine middle schools, urban in character, that found themselves in the bottom half of the state's API performance distribution. Within this band, the schools differed with respect to absolute performance and growth over time, but were as similar as possible with respect to social background factors. The study employed mixed methods and drew from multiple data sources: statistical analysis of teacher and student survey data, quantitative and qualitative analysis of classroom observations, ratings of student work samples, and interviews with administrators and teachers, as well as school background data. In total, our analysis is based on 317 teachers responding to a 340-item questionnaire, 4,148 students responding to a 50-item questionnaire, 270 observed lesson segments in English language arts, 390 pieces of student work, and 157 interviews. The study employs a structured multiple-cases design that allows for quantitative and qualitative cross-case comparisons (Miles & Huberman, 1994, pp. 172-238; Yin, 2003). It uses descriptive statistics and significance tests and applies the power of descriptive matrices to interpret quantitative and qualitative data from individual schools and various groupings of schools. The bulk of the data were collected in the 2004-05 school year.

Instruments

We strive to study practical relevance with rigor, and towards this end we have developed a number of robust research instruments for this study. All instruments were repeatedly field-tested. Factor, scale, interrater, and coding reliabilities were in most instances high and in a few instances acceptable. Many survey items and scales were validated in previous studies conducted by the authors and by other researchers in the field; some were specifically developed for this study. In the following section, we briefly describe the properties of our instruments and the ways they were administered. This will be discussed in more detail in a forthcoming technical report.

The *student questionnaire* consists of 50 items capturing the above-mentioned student perceptions of quality as well as questions about family background, awareness of accountability, and test-taking attitudes. It was piloted and subsequently administered to 4,148 seventh- and eighth-grade students. Students were sampled using a stratified random sampling technique in which we surveyed 50% of the classes in each curricular track. We adjusted for slight oversampling or undersampling with weights. The overall response rate was 96% (between 94% and 99% across the nine schools). See Appendix B for the statistical properties of major survey scales.

Classroom observations were conducted with the help of an observation instrument that we developed by adapting two previously validated instruments. We relied on the *Surveys of Enacted Curriculum* (Council of Chief State School Officers, Wisconsin Center for Education Research, & Learning Point Associates, 2003) and the Center for the Improvement of Early Reading Achievement School Change Observation scheme (Taylor, 2003). The protocol evolved into two parts, which are used simultaneously to allow observers to capture classroom teaching in its basic dimensions, but also pick up on more cognitively complex teaching occurrences. In total, we observed 90 English language arts lessons and classified 270 snapshots across the nine schools. Two observers who were trained extensively in a pilot school observed almost all lessons. An average of 20 decisions or ratings per observation was expected from observers. Interrater agreement ranged from 77% to 94%. Classrooms were sampled using a random sampling technique in which two researchers observed 50% of the seventh- and eighth-grade classes in each curricular track. Throughout each lesson, we rated three 5-minute snapshots spaced evenly throughout observation. The classroom observations were followed by a post-observation interview in which we tried to ascertain how teachers had approached planning and whether the observed lesson was tied to possible strategies of instructional improvement. Finally, we wrote a descriptive summary of each lesson according to a specified observation guide. The main observation points relevant for this report can be gleaned from the tables in the sections Quality of Students' Educational Experience/Significance Tests, and Organizational Characteristics/Blind Ratings, below.

Student writing samples were collected from English language arts classes in each school. As with the student questionnaire, we sampled writing using a stratified random sampling technique in which we selected 50% of the classes in

each curricular track. Within each class, we requested three pieces of writing: one high-quality, one medium-quality, and one low-quality exemplar. We collected 390 pieces of writing from 130 classes. As with the student questionnaire data, we adjusted for slight oversampling or undersampling with weights. These writing samples were rated with the help of four writing rubrics that we adapted from Newmann, Secada, and Wehlage (1995). The samples were rated by two independent raters without knowledge of school identities or performance status. After extensive training, an interrater agreement of 90% on the 20% of the sample overlapping between the two raters was achieved.

The *teacher questionnaire* consisted of more than 180 individual response items designed to collect information on teachers' perceptions of accountability, leadership, organizational strength, motivation, efficacy, school program, and change strategy, as well as teacher background data. Items and scales come from a variety of sources (Consortium on Chicago School Research, 2003b; McLaughlin & Talbert, 1993; Mintrop, 2004; SRI International, Policy Studies Associates, & Consortium for Policy Research in Education, 2003). We piloted about one third of the items or scales, primarily the ones we developed for this study. Several items and scales were field-tested repeatedly until sufficient validity and reliability could be established. Major scales with their statistical properties are listed in Appendix C.

The teacher questionnaire was administered to all teachers in the nine schools. Overall response rate was 83%, ranging from 67% for School I to 94% for School E. To reduce response time for teachers, we created two forms with the bulk of the items overlapping between both forms. 151 teachers responded to form A and 166 teachers to form B.

We conducted 157 interviews with administrators, classroom teachers, and teachers on special assignment using two basic protocols. In the first round of interviews, we concentrated on leadership, organizational culture, and accountability; in the second round, we inquired about instructional program and change strategies. Interview data are not used in depth for this report. In addition, we collected data on the schools' demographic background characteristics, school conditions, and inputs.

Case Selection

Initially, our study was designed as a comparison between five high-performing and five low-performing urban middle schools that ideally began from a

similar baseline API performance in 1999, proceeded with decidedly different growth patterns, and ended up on significantly different API levels at time of data collection, while at the same time being as similar as possible demographically. The criteria were, thus:

Demographic similarity;

Below-state-average baseline (1999) performance (1st to 4th decile);

Similar starting API in 1999 at the inception of the system;

Significant difference in absolute performance levels at time of data collection;

Contrasting high or low growth on API over a period of 4 years.

With these criteria we were able to select schools from typical urban environments and control for demographic background variables.

Recruitment of schools was a challenging process and took longer than expected. In the end, we settled for nine schools. Due to intensifying accountability pressures, a large number of low-API schools declined to participate, whereas it was much easier to recruit top performers. Regardless of API status, however, insecurity regarding one's performance status held sway across the whole nine-school sample since all schools were in below-average performance ranks, and many faced multiple and uncertain sanctions due to federal and state regulations.

Initially, we identified schools with exceptionally high and low growth on the API by predicting annual API achievement based on school background characteristics and calculating residual gains for each year. Thus, we made our case selection from groups of schools that grew well above or well below average on the API over a period of 4 years from 1999 to 2003, controlled for school characteristics.

We used data from the California Basic Educational Data System (California Department of Education, 2005), which contains annual API score data as well as information on school characteristics of 8,970 schools in California. We focused on only low-performing schools that ranked below the 50th percentile on their 1999-2000 API scores. Only middle schools or junior high schools that had a complete record of 4 years of API scores and demographic information were selected. We predicted API scores based on the School Characteristics Index (SCI), which is a composite index of the demographic characteristics (i.e., percentage of pupils with free or reduced price lunch participation, percentage of English Language Learners, student ethnic background, student mobility) and a proxy for school capacity (i.e.,

percentage of teachers with full credentials). The SCI is a variable contained in the state database. Since student populations and the basis for API changed over the years—for example, shifting from norm-referenced to standards tests—we could not simply sum API growth over time (although this is just what lay practitioners in the state do all the time), but instead calculated gains and residuals year to year.[3] We subsequently ranked schools according to growth residuals over time and identified schools in the top and bottom quartiles. These were the groups from which we made our case selections.

Schools we selected from the high and low ends of the performance spectrum had a similar baseline API in 1999: at least 60% of their students from disadvantaged minority populations (African American and Hispanic students); high poverty rates as indicated by at least 50% of free or reduced price lunch participation (FRPL); at least 20% of students with limited English proficiency; and an urbanicity score of at least 3 (= urban fringe). We excluded schools with total enrollment exceeding 2,500 students, charter schools, magnet schools, and year-round schools. The latter restriction cut out large numbers of schools in Southern California's low-performing districts, but for ease of matching school conditions the limitation was necessary.

Characteristics of the Nine-Case Selection

Table 1 shows the nine schools that chose to participate in the study; four of the schools are classified as low performing and five as high performing. During recruitment it became clear to us that schools with the most challenging conditions avoided participation in the study. Often swamped with audits and inspections, they felt they could neither spare the time nor benefit from one more external review. As a result, our four lower-end performers are biased towards those types of schools that "felt better than they appeared," and indeed all of the nine schools were clean places, pleasant to visit, and did not fit the stereotype of an "out of control," failing school, sometimes espoused by the media.

As Table 1 illuminates, schools in the low category differ from those in the high category by having lower absolute API performance and lower growth from 1999 to 2005, the last year we collected data. Although the distance between top-performers in the low category and bottom-performers in the high category diminished over time, most differ by more than half a standard deviation from the nine-school mean in absolute API performance. Overall, mean API for the high group is 660, for the

[3] We want to thank Professor Yeow Meng Thum for his assistance in calculating this growth model.

Table 1

Academic Performance Scores of the Nine Selected School Cases

	Low				High				
	F	D	I	C	H	G	A	E	B
1999 API	478	503	478	481	442	521	489	523	445
2005 API	573	573	598	604	642	653	653	670	683
Score difference	95	70	120	123	200	132	164	147	238
Standard deviations from 2005 mean API[a]	–1.3	–1.3	–0.7	–0.6	0.4	0.6	0.6	1.0	1.3
2000 State rank	2	2	2	1	1	2	2	2	1
2004 State rank[b]	1	1	1	1	2	3	3	3	4

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc.

[a] $M = 628$; $SD = 41.5$. The mean is calculated as the unweighted average of the nine schools' API scores and is slightly biased. The unbiased mean of the high and low groups is 624. Significant differences of means were tested using the Mann-Whitney test ($z = -2.47$, $p = 0.0135$).

[b] The 2004 rank was the last available score at the time of data collection.

low group 587, a 73-point difference that is statistically significant at the .01 level. Movement in state ranks corroborates these group differences. All nine schools started in 2000 either in the lowest or second lowest API decile. Four years later, the four schools classified as low either declined or remained in the lowest rank, whereas by contrast, the five schools classified as high moved up at least one decile; one school moved up three deciles.

While the two groups differed in API performance, both in absolute and relative terms, they were quite similar demographically. None of the school background indicators displayed in Table 2 show statistically significant differences across groups, though they differ within groups. Three of the four schools in the low group tended to be economically more challenged as indicated by higher FRPL participation, whereas schools in the high group had higher proportions of English learners. Two schools (Schools I and C) had relatively lower proportions of African American and Hispanic students, but a high proportion of Hmong students.

To explore school context conditions with higher grain size, we inquired about student and teacher perceptions of family background and support for education. Three of the four scales show statistically significant differences across groups (see Table 3), but schools classified as low were only more challenged in the area of parental support (due to the low ratings of one school), whereas by contrast, the

Table 2

Demographic Characteristics of the Nine Selected School Cases, 2004-05[a,b]

	Low				High				
	F	D	I	C	H	G	A	E	B
Enrollment	866	1,100	1,031	991	1,818	705	1,628	780	868
African American (%)	3	4	9	12	0	1	5	6	1
Hispanic (%)	88	84	56	59	97	59	75	81	93
English Learners (%)	29	22	39	26	44	31	43	18	28
Free/Reduced lunch (%)	97	59	100	100	77	85	83	69	78
Parent education[c]	1.81	2.13	2.09	2.25	1.81	2.02	2.09	2.18	2.03

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc.

[a] Source: California Department of Education (2005).

[b] All means are statistically insignificant between high and low groups using the Mann Whitney test.

[c] 1 = Not a high school graduate; 5 = Graduate school.

Table 3

Teacher and Student Perceptions of Family Background—Mean Response

	Low				High				
	F	D	I	C	H	G	A	E	B
Teacher-reported parental support* (range: 7–32)	13.9	17.0	17.9	17.7	18.5	20.1	19.1	18.6	19.1
Student-reported familial support (range: 6–24)	16.8	18.2	16.9	17.7	16.9	17.3	17.7	17.9	17.0
Student-reported possession of cultural goods** (range: 1=none, 4=all)	2.2	2.2	2.1	2.1	2.1	2.1	2.1	2.0	2.0
Student-reported frequency of non-English home language** (range: 1=never, 4=always)	3.0	2.7	2.7	2.9	3.3	3.2	3.2	3.0	3.4

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc.

*$p < .05$. **$p < .01$.

higher performing schools were more challenged in the area of language and possession of cultural goods. Thus, without ignoring the potentially higher challenge of poverty in some schools in the low group, as indicated by high percentages of FRPL participation, the groups overall seem fairly well matched demographically. In conclusion, our nine-school sample is a suitable case selection

for our intended analyses. It shows substantial differences in student achievement as expressed by API, but is sufficiently similar demographically.

Testing the Model of Practical Relevance

To recount, our model of practical relevance refers to five areas: achievement, similar educational challenge, educational quality, organizational effectiveness, and accountability. We asserted earlier that the practical relevance of accountability systems would be enhanced to the degree that the prime system performance indicators line up with other indicators of educational quality, effective schools characteristics, and engagement with accountability. Having convinced ourselves that the first two conditions are in place, we now proceed to test the latter three conditions.

Two analytical steps make sense. First, we look at the relationship between absolute API-indicated performance and other educational quality indicators (i.e., quality of students' educational experience), keeping in mind that higher student achievement as measured by the indicator should be reflected in the way students experience their school. Second, we look at the relationship between API-growth differentials and organizational characteristics in conjunction with accountability scales, keeping in mind that higher growth in API over time should be reflected in a better functioning organization and a more productive response to accountability.

For each of these analytical steps we apply three procedures. First, we conduct blind ratings of schools with the help of case-ordered descriptive meta-matrices (Miles & Huberman, 1994, pp. 190-192). This method retains the record of each individual school for decision making. Second, we conduct statistical significance tests across various school groupings. Third, we investigate configurations unique to individual schools using both quantitative and qualitative data. A more in-depth look at these individual configurations has to be left for another report.

The Quality of Students' Educational Experience

Is absolute performance level as measured by the state indicator matched by other quality criteria that circumscribe students' educational experience? As a reminder, our high group differs from our low group by a mean 73 API points, and the difference between our top school and bottom school is 110 API points. These differences are not trivial given that it took our low-growth schools 6 years to make a gain of 70 to 100 points.

Blind ratings. To avoid bias in our ratings, we concealed from ourselves all information that could identify the schools and their performance status. Since the research team had visited the schools numerous times in the course of 1 or 2 school years, we chose not to use any interview data for this first rating since the likelihood of recognition would have been too high. We then constructed, based on measures from student questionnaire data, classroom observations, and writing sample ratings, a matrix that indicated whether a given school was on, above, or below average in a given measure.

We marked the continuous student perception variables by comparing scale means and standard deviations across the nine schools. We assigned a zero to school means that fell within one standard deviation of the nine-school mean, a plus (+) or minus (–) to means that fell more than one full standard deviation above or below the mean, and an asterisk (*) to denote borderline cases, alerting us to possible classification uncertainties. We proceeded similarly with the continuous writing scores, except that we excluded one negative outlier school score from the calculation of the across school mean. For the categorical variables, we divided the range of scores into four equal intervals and assigned plus (+) or minus (–) marks to scores in the top and bottom intervals and zeros for scores in the middle, allowing an asterisk (*) for borderline cases.

We then represented these data in a matrix (suppressing 0 marks), which we called a school profile. This profile was fairly unbiased in that student perceptions were not identifiable by school, writing samples had been rated blindly in the first place, and classroom observations were validated by high interrater agreement. We then tried to predict an individual school's performance status by looking for consistent patterns across our multiple indicators of quality. This involved judgments not unlike ones that would have been made by our imagined traveling practitioners. Two raters with long experience in schools studied each school profile and then judged whether a school was a likely high or low performer or whether the case was undecided. For the "undecideds," we allowed raters to indicate at least a direction up or down if they so chose. As a decision rule, we elected that if we could identify at least half the schools correctly (in the high group, three out of five) and the rest at least as undecided without interrater disagreement, we had "validated" API performance status.

Table 4 displays our decision-making matrix and ratings with interrater agreements, matched with the (previously concealed) absolute API scores at the time

Table 4

School Profiles: Quality of Students' Educational Experience

	Low				High				
	F	D	I	C	H	G	A	E[a]	B
2005 API Score	573	573	598	604	642	653	653	670	683
Academic engagement	–		+				+		
Academic press	–								
Teacher care	–		+				+		
Peer collaboration	–			–					+
Safety			–	–		+	+		
Suspension rate (lower: +)			–	–	+	+	+		+
Non-instructional time (lower: +)	–	+		–	–	+	+	+	
Time on task	–			+	–	+	+		
Student engagement	–	–	+	–	–		–	–	+
Positive teacher tone	–	+	–	–	+	–	–		+
Proactive instruction		+		–		–		–	+
Cognitive complexity		+	+			–			+
Writing score				–	+	+	–		
Blind summary ratings	↓	↑	0	↓	0	0	0↑	0/↓	↑

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc. ↑ = Possibly high; ↓ = Possibly low; 0 = Undecided.

[a]Interrater disagreement.

of data collection. As can be seen, it was not possible to correctly and reliably classify a sufficient number of schools in the appropriate performance status group. Negative ratings for quality of students' educational experience were more frequent in the low group, but one school (School D) was rated as high by both raters. Schools in the high performance status group were not consistently rated better in terms of student perceptions, classroom teaching, and quality of student writing, standing together, with the exception of the top API school (School B).

Significance tests. Our purpose in using statistical significance tests in this study was not to arrive at generalizable findings, but to make ourselves more independent of the subjective rater judgments. We conducted significance tests for all our student perception scales. We checked for differences in means across our two performance groups by conducting weighted survey regression analyses. These have the advantage of counting each school's means with the same weight (Kish, 1965). None of the weighted regressions turned up statistically significant

differences of means. Table 5 shows that as far as student perceptions are concerned, the nine schools were all very similar. Students on the whole tended to feel neutral to mildly positive about their schools. These were not students who felt particularly excited about the educational experience provided to them, at any of the nine schools, regardless of the schools' API scores. According to their own perceptions, students on the whole were neither more challenged nor more academically engaged in the high-API schools.

As to classroom observation measures, Table 5 suggests that no consistent patterns obtain for quality of lessons across our high- and low-performance groups (consult Appendix D for information on how we composed classroom observation measures). Indeed, these measures are all statistically insignificant using the Mann Whitney test. Thus, neither blind ratings, nor significance tests, nor descriptive analysis rendered a clear and consistent pattern of higher quality of students' educational experience in high-API schools. It should be noted, however, that suspension rates (and perhaps non-instructional time) tended to be higher in the low group, and writing quality was slightly higher in four of the five schools classified as high.

Individual schools. While consistency in educational quality between our two API status groups is difficult to establish, the top API school (School B) stands out with positive marks in many measures and no negative marks. Both blind raters also agreed that School B provided a high-quality educational experience to its students, relative to the other schools in our case selection. But School B in the high-API group is not the only one that was so rated. Indeed low-API School D was also rated as high by the blind raters, primarily because of a higher frequency of lessons (strictly speaking, lesson snapshots) in which the teacher tone was positive, the teaching was proactive, and learning activities went beyond mere recall. Although "time on instruction" measures are on par with the top API school (School B), School D lacked the sense of student engagement that pervaded the lessons observed in School B, as classroom observation data confirm. These differences notwithstanding, it is noteworthy that instructional quality (but not student achievement as measured by standardized tests) at least in the English language arts classes was remarkably similar in the two schools irrespective of a formidable 110 point difference in API at the time of data collection.

Two schools in the low-API group, Schools I and C, stand out with exceptionally high suspension rates. This, in combination with somewhat lower

Table 5

Measures of Quality of Students' Educational Experience

	Low				High				
	F	D	I	C	H	G	A	E	B
Student Perceptions (Mean)									
Academic engagement (Scale midpoint: 17.5)	17.9	18.7	19.7	18.0	18.0	18.5	19.4	18.3	18.3
Academic press (Scale midpoint: 10)	12.3	13.3	13.2	13.0	13.0	13.2	13.3	13.3	13.2
Teacher care (Scale midpoint: 12.5)	13.5	14.5	14.7	13.8	14.0	14.3	14.8	14.1	14.1
Peer collaboration (Scale midpoint: 10)	12.1	12.7	12.8	12.1	12.2	12.5	12.7	12.7	12.8
Safety (Scale midpoint: 7.5)	9.1	9.4	8.8	8.7	9.1	9.7	9.5	9.3	9.1
Suspension rate (Percent)	44	32	59	48	9	21	21	25	13
Enacted Curriculum (Percent)									
Non-instructional time	14	3	6	15	13	0	0	0	6
Time on task	82	89	87	96	80	100	100	92	93
Student engagement	6	9	17	4	7	9	4	8	20
Positive teacher tone	50	84	59	59	81	57	50	70	80
Proactive instruction	27	60	47	15	36	25	48	26	50
Cognitive complexity	21	51	44	33	29	12	40	29	53
Mean writing score	7.3	7.2	7.0	3.9	7.9	8.0	6.4	7.5	7.6

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc.

student safety ratings, may hint at a higher disciplinary burden compared to schools in the high-API group (particularly Schools G and A). Whether this is due to ineffectiveness or social context is not entirely clear, but extremely high FRPL participation rates in these two schools (see Table 2) point in the direction of social context differences that create challenges encountered less in the higher API schools. Low-API School F may be similarly challenged given its lunch participation rate and teachers' exceptionally low ratings of parental support relative to the other eight schools (see Table 3). Qualitative data confirm a much higher concern for, and effort expended on, safety and order in three of the four low-API schools. On the other hand, two high-API schools, Schools A and G , stand out as very safe and orderly schools, though they also are highly impacted by poverty as indicated by FRPL rates of about 85%.

Conclusion. We surmised earlier that an accountability system increases its practical relevance to the degree that its prime performance indicators are clearly and consistently associated with quality in students' educational experiences. We have found that, in the case of nine California schools, information from system indicators and patterns of educational quality as measured by the conventional criteria of this study are not closely matched. Our traveling practitioners would indeed have a hard time distinguishing indicated high-performing from low-performing schools by observing English teachers and talking to students, but perhaps gain a slightly better idea by reading students' writing. Without knowing actual test results, they may lump schools from different performance groups together and may at least question differential educational challenges that are not adequately represented by official statistics.

Organizational Characteristics

Is high or low growth as measured by the prime state indicator matched by organizational effectiveness criteria that circumscribe adults' interactions with each other and responses to the accountability system? Exploring the practical relevance of accountability systems in the area of organizational effectiveness requires us to look at growth over time in the performance indicator, rather than absolute performance levels in a given year because it is the process of improvement that we associate with the superior quality of adult interaction. Schools with high absolute performance could just as well be "cruising schools" (Stoll & Fink, 1998), which manage to maintain their status by capitalizing on environmental advantages without exerting greater internal strength. With this in mind, we selected into our high and low performance status groups only schools that not only had substantially different absolute API levels, but also arrived at these levels due to either exceptionally high or low growth over time.

Our original sampling and classification was based on 4 years' growth on the API between 1999 and 2003. Based on those years, schools in our two performance status groups differed nicely on both absolute and relative measures. The two top- and bottom-growth schools differed by 100 API points (168 points by 2005), while the marginal difference between the two groups was about 40 points, a year's increase on the high end for most of our selected schools.[4] This difference

[4]Our language is somewhat vague here because we want to capture the intuitive approach of practitioners who tend to add API points over time and be mindful of the statistical properties of the API that do not allow straightforward additions.

diminished to a mere 9 points by 2005, making the original "high" and "low" group distinctions less pertinent than was intended with our erstwhile case selection. On the other hand, schools in our low group had all either stagnated or declined in their state rank, that is, API performance decile, while all of our schools in the high group had grown at least by one rank.

Blind ratings. As we did for the educational quality ratings, for the School Profile (Table 6), to avoid bias, we concealed from ourselves all information that could identify the schools and their performance status and constructed a matrix based on measures from teacher questionnaire data, for the time being excluding interview data. When marking continuous variables on the teacher questionnaire, we assigned a zero to school means that fell within .1 point of the nine-school mean (suppressed in the school profile); one plus (+) or minus (–) to means that fell less than one standard deviation above or below the mean; two pluses (++) or minuses (––) to means that fell more than one full standard deviation above or below the mean; and an asterisk (*) to denote borderline cases.

Rating schools' growth based on this profile seemed more difficult than rating the "educational quality" profile since so many more measures had to be considered. Blind raters basically searched for a preponderance of pluses over minuses, while taking various scenarios into account (e.g., strong accountability response due to high pressure/high meaningfulness; strong controlling/collegial leadership; strong organizational culture due to cohesion/openness; strong motivation; high sense of efficacy; high degree of rationality in change strategy).

Table 7 shows the blind ratings juxtaposed with (previously concealed) API-growth differences. Despite the many more measures to be considered, raters only disagreed on two schools. Our decision-making rule for classification was the same as before: We needed to at least classify half of the schools (three out of five in the "high" group) correctly and without interrater disagreement.

We first tested the original high/low group distinction, which is based on absolute API performance in 2005, calculated residual growth from 1999 to 2003 (not displayed here), movement in state rank, and overall API differences summed over 6 years. Clearly the raters were unable to classify the schools correctly. More schools in the high group were classified as less effective than more effective; and two schools in the low group were rated as more effective.

Table 6

School Profile: Organizational Effectiveness and Response to Accountability

	Low				High				
	F	D	I	C	H	G	A	E	B
Accountability									
Goal importance	+	−−	+	+		−−	−	−	++
External validation	+	−	+	+		−−	+	−	++
Authoritativeness	+			*+	+	−−	−	−	++
Threat	+	+	−	−	−	−	−	−	−
Pressure	+	+	−−	−−		+	−	+	
Focus	−−	−	+	+	−	−	+	−	++
Diagnostics	−−	+	+	+	−	−−	+	+	++
Validity	−−	−−	++	+	+	−	+	−	++
Fairness	−−	−	+	+			++	−	++
Realism	−−	+	+	−	−	−	−	+	++
Raised expectations	+	−−	+	+	−	−	−	−	++
Goal integrity	−	−		++	*−	−		*−	++
Student-reported test importance	−		+	−		+	+	+	+
Leadership									
Urgency	++	*−		+		−	−	−−	*+
Principal support	+	+	+	++	−	−	−	−−	+
Principal control	++	−−	−	++	−−	+		−−	+
School management	+	−	−	++		+		−−	+
Open communication		+	+	++	−	−	−	−−	+
Autonomy	+	*+		*+		*−	−	−−	*+
Instructional leadership	*+	−	−	++	−	+	−	−−	+
Moral leadership	+		−	++	−	+	−	−−	++
Faculty Culture									
Collegiality	+		−	++	−	−	+	−−	++
Pulling together		−	+	++	−	−	+	−−	++
Norms of performance	+	−	−	++	−	+	−	−−	++
Learning orientation	+	+	−	+	−−		+	−−	++
Motivation									
Involvement									
Hard work	−	+	*−	−−	−	+	−	++	++
Commitment to stay	−			+	+	−			
Morale/Improvement expectations	−	−	+	++	−−		++	−−	++
Satisfaction	+	−		++			+	−−	++

(*table continues*)

Table 6 (continued)

	Low				High				
	F	D	I	C	H	G	A	E	B
Efficacy and Qualifications									
Instructional efficacy									
Test-related efficacy									
Colleagues' skills	+	–		+	–			––	++
Preparedness	+	–		+					
Total years teaching	–	––	++	++	+	–	+	–	–
Degree	–		–	–			–	+	–
Full certification	+	+				+	–	+	
Change Strategies									
Program coherence	+	––	+	++	–		+	––	+
Strategic orientation	+	–		++	–	–	+	––	++
Planning	+					+		+	–
Data usage									
District operational system	+	–		+	+	++	+	––	––
District instructional system	+	––	+	+		++	–	–	––

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc. Zero = school means fell within .1 point of the nine-school mean (suppressed in school profile); one plus (+) or minus (–) = means fell less than one standard deviation above or below the mean; two pluses (++) or minuses (––) = means fell more than one full standard deviation above or below the mean; an asterisk (*) = borderline cases.

Table 7

Growth Ratings Based on Organizational Effectiveness and Accountability Response

	Original "Low"				Original "High"				
	F[a]	D	I	C	H	G	A	E[a]	B
1999-2005 API difference	95	70	120	123	200	132	164	147	238
State rank 2000-2004	–1	–1	–1	0	+1	+1	+1	+1	+3
2003-2005 API difference	36	–4	65	56	47	–4	37	36	78
Blind summary ratings	0/↑	↓	0	↑	↓	0↓	0	↓/0	↑

Note. A, B, C, D, E, F, G, H, and I = School A, School B, etc. ↑ = Possibly high; ↓ = Possibly low; 0 = Undecided.

[a]Interrater disagreement.

Significance tests. To double-check our subjective ratings, we conducted significance tests for all our teacher perception scales. We checked for differences in means across the nine schools by Analysis of Variance (ANOVA), and across our two performance groups by conducting weighted survey regression analyses. In

keeping with the multiple-case design of the study, we were not interested in the statistical significance of single measures. Instead, practical relevance of the accountability system was bolstered only if growth on the performance indicator "fit" the earlier hypothesized models or scenarios of effectiveness. This meant that multiple measures had to show up as significantly different, standing together as a pattern that could be corroborated through subsequent analysis of more in-depth information gleaned from the interview data.

A straightforward comparison of the nine schools, five originally classified as high and four as low, did not reveal such a pattern. Quite the opposite. With one exception, none of the school effectiveness and accountability measures listed in the School Profile (Table 6) are significantly different. Reducing the comparison from nine to five schools (three top API-growth schools versus two bottom API-growth schools over 6 years), thus increasing the marginal difference between the two groups, makes three accountability-related measures (focus, validity, fairness) significant at $p < .05$, but this is still far from the kind of more encompassing patterns that could establish practical relevance. Thus, for 6-year API score differences (1999-2005), absolute performance level on the API in 2005 (time of data collection), and movement in state rank from 2000 to 2004, our measures of organizational effectiveness and accountability response do not work, neither with a pressure and control scenario, nor with a meaningfulness and organizational learning scenario. Blind ratings and statistical significance confirm each other.

Searching for a better fit. On the other hand, an even cursory look at the School Profile matrix tells us that some schools are more effective than others according to our measures; and further statistical exploration through analysis of variance across the nine schools (not displayed here) corroborates that the means of almost all of the organizational effectiveness and accountability measures differ statistically across the nine schools. School B and School C stand out as the two cases that were rated unambiguously by the two blind raters as high-growth. But only one of the schools (School B) was grouped in the original high group, and one was originally thought of as low performing. Was this further indication of the performance indicator's irrelevance in terms of organizational effectiveness or was there a connection yet to be uncovered?

Trying out a number of combinations, we found that the two schools grew rapidly in the last 2 years while some previously high-growth schools declined and low-growth schools soared. Things had apparently shifted during the year and a

half of data collection. Instability of growth trajectories and effectiveness status has been noted repeatedly (Elmore, 2004; Gray, 2001; Teddlie & Stringfield, 1993), and our schools were no exception in this regard. Gray (2001) wondered "whether five years is a life time or a brief moment in a school's natural history" (p. 1). For some of the nine schools in this study, 6 years, the time span we inquired about, is two life times, given leadership changes, teacher turnover, student mobility (particularly in middle schools), and the fluid social composition of California immigrant communities. And as a consequence, student and teacher perceptions about their school in one single year are rarely good for much longer.

When tested for change over merely 2 years, some organizational characteristics show systematic relationships to API growth. Table 8 displays the results of weighted regressions. The independent variable is dichotomous and consists of two performance groups: three schools (Schools B, C, and I) that posted the highest API score differences from 2003 to 2005 and five schools (Schools A, D, E, F, G) that posted substantially lower API score differences, leaving out School H to provide for sufficient distance between high and low API-growth schools (see Table 7). The dependent variables are continuous perception scales. Only scales that showed statistical significance are displayed.

To begin with, one should not overestimate the significance of these statistics. Simply adding API score differences over 2 years adds measurement error; and there is a certain element of arbitrariness in the groupings. Given the low number of cases, results can sway depending on what schools one decides to group together. The groups compared in Table 8 differ in terms of 2-year API growth, with the three high cases growing between 56 and 78 API points and the five low cases growing between –4 and 37 points. On the other hand, the results remain fairly similar when comparing more extremely composed groups, for example, the three top- and two bottom-growth schools, suggesting a more stable pattern.

Although the means for many of the scales are fairly close together and hover around scale midpoints, the significant measures, *together*, speak to a conspicuous pattern: Rather than generic characteristics of effectiveness (for example, strong leadership), it is a school's specific response to accountability that seems more central, especially the degree to which the system is internalized. One could surmise the following scenario from these data: Compared to similarly situated lower API-growth schools, schools that grew strongly over the last 2 years attach greater importance to accountability goals. They are more concerned about their external

Table 8

Organizational Characteristics of Higher and Lower API-Growth Schools, 2003-2005 (Survey Linear Regression)

		Estimated mean	
	Range	Lower 5	Higher 3
	Accountability		
Goal importance**	4–20	13.9	15.5
External validation**	3–15	9.5	11.1
Authoritativeness*	3–15	10.5	11.5
Pressure*	1–5	4.3	3.8
Focus***	3–15	9.6	11.3
Diagnostics*	5–25	15.0	17.7
Validity**	3–15	6.6	8.2
Raised expectations**	4–20	12.2	14.4
Goal integrity*	1–13	8.3	9.6
	Leadership and Faculty Culture		
Open communication*	4–20	12.4	14.6
Pulling together**	3–15	9.4	11.6
Morale/Improvement expectations*	1–4	3.1	3.4
Colleagues' skills*	3–15	11.2	12.0
Commitment*	1–3	2.4	2.5
	Change Strategies		
Program coherence*	4–20	11.2	13.8
Strategic orientation*	2–10	6.5	7.6

*$p < .05$. **$p < .01$. ***$p < .001$.

reputation and are more willing to accept the normative authority of the state to direct them. But by the same token, they regard the system as more meaningful: They see it as more useful and have fewer doubts about its rightfulness. They believe that accountability demands led them to raise expectations for themselves and their students. And they feel a better agreement between system demands, student needs and their own values. They sense to a higher degree that they pulled together as a faculty around accountability demands, but at the same time there is greater opportunity to have open discussions about the meaningfulness of accountability and chances to disagree. Expectation of improvement, trust in collective capacity, and commitment to stay are slightly higher. Teachers perceive their school as better organized with regard to instructional program and change strategies. They do not feel more pressured or personally threatened by sanctions

than their lower growth counterparts. We characterized this pattern as constructive engagement. Schools exhibiting constructive engagement do not embrace the accountability system with flying colors, but are willing to search for ways to positively engage with it and use it to move an improvement agenda forward.

Individual schools. A look at individual schools (see Table 6) clarifies, refines, but also questions the pattern derived from group comparisons. Two of the top three high API-growth schools (Schools B and C) conform most closely to the hypothesized model of organizational effectiveness and constructive engagement with accountability. Indications of meaningfulness and internalization of accountability were higher; principal leadership was stronger in managerial, collegial, instructional, and moral terms. Faculty culture was stronger and morale was up. But School I, a school similarly as high-growth as School C, does not fit this pattern. In the areas of principal leadership and faculty culture, the school looks like a less effective school, yet accountability was more internalized, the school pulled together in the face of accountability demands, and program coherence was more strongly in place. One gains a better understanding of the school through interview data. The principal carefully monitored instructional program coherence (meaning here the faithful implementation of the main language arts program according to district pacing guides) and basic social order, but beyond that played a generally supportive, but benignly distant role.

By contrast, School F, a lower API-growth school, was led by a strong principal who came to emphasize control, urgency, and the pressure of accountability to move his faculty forward. Teachers, on the other hand, tended to see the accountability system in negative terms and did not connect it to instructional practice to the same degree as teachers in School C. A defensive posture and confrontational attitude developed. Low morale set in, for which leadership efforts and the forceful monitoring of instructional program implementation could not compensate. Leadership strength was not sufficient by itself, this school seems to suggest, without some internalized acceptance of the accountability system as guidance in the area of instruction.

School D was the lowest growing school in the nine-school sample. Here, teachers saw their faculty as less cohesive and their principal as more open and supportive, but lacking in other aspects of leadership. The accountability system loomed as threat and high pressure, probably due to the school's recent decline in API. Neither principal nor district leaders seem to have communicated urgency. Yet,

a strong streak of opposition to the accountability system as incompatible with the school's philosophy of student-centeredness and professionalism pervaded this faculty more so than any other. Interestingly, both blind raters rated School D high in educational quality and low in organizational effectiveness.

School G was the other school in the selection with negative API development in the last 2 years. This school also had relatively lower engagement with accountability, but was higher on leadership. Here, the relatively new principal, strongly backed by district officials, exerted strong control and enforced norms of performance around discipline and time on task. But other aspects of instruction were relatively untouched.

A comparison between School I and School A confounds our model the most. Both schools are very similar on almost all measures, yet one grew by 30 more points in the last 2 years. They both engaged with the accountability system constructively and perceived their principal as relatively weaker. Interview data revealed that the principal at School A was seen as less of a presence in instructional affairs compared to the higher growth School I, yet the faculty had a strong collective tradition, preceding the current principal's tenure. Thus, uncertainty remains.

Conclusion. We surmised earlier that an accountability system increases its practical relevance to the degree that its prime performance indicators are clearly and consistently associated with organizational characteristics of effectiveness and engagement with accountability. In the case of nine California schools, growth status as measured by system indicators and effectiveness as measured by the conventional criteria of this study are matched to a degree. Our fictitious traveling practitioners could learn some valuable lessons if they selected the right time frame. If they selected schools based on absolute API or growth over a longer time, no stable and consistent contrasts between high- and low-performing schools' organizational characteristics could be discovered. If they used a shorter time frame for the selection of schools to be visited, they could learn that leadership, as a combination of management and learning facilitation, a cohesive faculty culture with strong norms of performance, and constructive engagement with the accountability system, coupled with implementation of a structured language arts program, were more developed in schools that experienced recent growth on the state performance indicator. But they would also find schools that grew without this exceptional leadership and faculty culture and schools that had stronger principal

leadership that did not grow. The latter, however, seem to rely more on control and amplification of system pressure and threats. A closer look might then reveal the essential force, absent in these lower growth schools: a stronger belief in the meaningfulness of the accountability system coupled with some basic leadership support and efforts to focus on a coherent and aligned instructional program at least in the area of literacy. But our travelers, not unlike these researchers, would also be confounded by schools that do not seem to fit this pattern.

Synopsis and Discussion

Whether or not, or to what degree, we attest practical relevance to an accountability system depends on a robust relationship between indicated performance status and clear and consistent patterns in the three dimensions of educational quality, organizational effectiveness, and accountability. We have found that the system's practical relevance for school improvement is limited, but not without merit, given our definition of the term and our selection of nine cases.

To begin with, indicated absolute performance level could not be systematically linked to any of the aforementioned three dimensions. At the individual case level, only the highest performing school in the nine-school selection stands out. It is the only school that appears strong in all three dimensions and the one school that is an outlier even within the "high" group and indeed in the statewide sample for our demographic profile. To the degree that one can learn from outliers, this school bolsters the case for the system's practical relevance.

But, as we saw, not even this outlier school, let alone the other high-API schools, can be clearly distinguished from a much lower performing school along educational quality criteria. Whereas the lack of systematic connection between absolute performance and measures in the organizational and accountability dimensions can be explained by the coasting phenomenon (i.e., schools could already be in a stagnation or even decline pattern while still benefiting from earlier growth), a lack of systematic connection between API performance and educational quality is more difficult to accept and seems to be grounds to question the practical relevance of the accountability system.

For growth, the picture looks better. Confounding individual cases notwithstanding, indicated short-term growth was connected to consistent patterns of organizational effectiveness and accountability. The highest degree of consistency

was obtained for the accountability dimension. How can this configuration be interpreted, and what does it mean in terms of the system's practical relevance?

At first blush, degrees of differential consistency across the three dimensions can be conceived according to proximity to the indicator. Attitudes to accountability are more directly proximate to movement in standardized tests than organizational culture in general or teacher behavior and student perceptions of schools. Faculties that attach greater importance to accountability goals are presumably also more likely to pay careful attention to the system's standardized tests and ways to improve on them. After all, it was not foremost the coercive aspect of accountability that held sway in the higher growth schools, but a greater sense of meaningfulness that may help internalize accountability into the instructional core. These accountability attitudes should not be regarded as stable cultural patterns that coagulate in schools with a past of indicated high growth and present high performance. Rather, they seem to appear in our cases during upswings, perhaps to be lost again in subsequent years so that they do not show up as consistent associations with high absolute performance at a point in time.

A more positive attitude towards accountability could theoretically be cause, coincidence, or result of higher growth. Performance success and positive attitudes towards the judging authority often go hand in hand and can mutually reinforce each other. Sense of pressure and threat seems to have developed in School D as a result of the school's recent performance record, independently of administrative pressure, but more positive engagement with accountability was not encountered in schools that had posted solid growth over 6 years, as could have been expected. Instead, it occurred in three schools where principals had actively forged a consensus around accountability. Two of the three schools (Schools C and I) had very recently gone on the upswing, yet were still classified by the system as very low performing and on the verge of major corrective action. Teachers in these schools were still insecure about their future prospects, but nevertheless more positive about accountability.

It is conceivable that when schools grew on the API they also systematically improved on other criteria of educational quality. This cannot be directly tested in this study due to the nonlongitudinal design. But it is not very plausible, given that we can find so little consistent association between high absolute API performance and (other) educational quality measures. A more plausible explanation seems to be that teachers in high-growth schools connect to the accountability system more

strongly and do something that results in higher standardized test scores, but that something is not necessarily a marked change in their teaching practice, nor does it strongly influence students' perceptions of their educational experience. In other words, what teachers do to increase performance on the state tests may not necessarily translate into higher academic engagement of their students, better teaching, or more learning complexity, nor does it seem to even influence time on task or academic press in a consistent manner. Yet, in the successful schools, teachers do improve student achievement as measured by standardized tests. How?

Our observations suggest that teachers in these schools have committed to a highly focused coverage of standards-aligned materials within highly structured literacy and language arts programs that are taught in differentiated learning groups. This approach, the study seems to suggest, does not necessarily translate into better teaching or a richer educational experience for students, though it may have had positive consequences for quality of students' writing.

A comparison between the top API School (School B) and the bottom API School (School D) illustrates what we mean. Both schools have been described in previous sections. Compared to the top school, the bottom school was less effective organizationally. It rejected the accountability system and the highly structured language arts and remedial literacy programs that aligned with the system. Our quantitative measures register this as below-average (for the nine case sample) *meaningfulness of accountability* and *instructional coherence* ratings. But this did not mean that teachers in this school taught any worse or provided a poorer learning environment for students (i.e., student perceptions and data from lesson observations were very similar, though writing quality was lower). The top API school by contrast was enthusiastic about the accountability system and had decided to focus its energy on curriculum alignment and structured programs. The majority of below-grade-level students were taught for the majority of their learning time in remedial literacy programs. Social studies and science had been de-departmentalized in this middle school and folded into the teaching of the literacy programs. The programs were implemented well, reflected in an above-average proportion of engaging lessons. Doubts were openly aired in this faculty about the adequacy of science and social studies instruction, but for the time being, a collective commitment had been made to focus.

Eight of the nine schools, the above-mentioned School D being the exception, follow in the footsteps of School B. But they did not implement their standards-

aligned and structured programs with nearly as much enthusiasm and to the exclusion of other subject matter, although in seven of the nine schools electives had been abandoned in favor of language arts or moved to the realm of voluntary after-school activities. Some schools implemented the programs with a heavy emphasis on monitoring and failed to generate commitment by internalizing accountability. In others, the programs were in place, but implemented with a more casual attitude. In neither case can program implementation be connected to an overall higher quality of students' educational experience.

In summary, we initially stated that practical relevance of the accountability system for school improvement would be high if our fictitious practitioners could learn from their travels across a spectrum of schools that contrasted on indicated performance but were similar in their educational challenge. We surmised that lessons should be learned around schools' response to accountability, organizational effectiveness, and educational quality. Traversing the nine schools that we studied here, our travelers would learn that schools that grew on the performance indicator tended to generate internal commitment for the accountability system. They eschewed the coercive aspects of accountability, maintained a climate of open communication, and considered the system as an impetus for raising expectations and work standards. On the instructional side, this commitment translated into the forceful implementation of structured language arts and literacy programs that were aligned with the accountability system. If our travelers expected to encounter visible signs of an overall higher quality of students' educational experience in the high-performing schools, they would be disappointed. Rather they would have to settle on a much narrower definition of quality that homes in on attitudes and behaviors that are quite proximate to the effective acquisition of standards-aligned and test-relevant knowledge but go beyond mere teaching to the test, as the quality of student writing seems to suggest. Whether they would settle on such a narrower definition would depend on the meaningfulness of state standards and tests and their own educational inclinations. But in many respects, schools that are classified as high performers turn out to be questionable role models for their less fortunate counterparts.

Implications for System Design

Though we have never found a theoretical justification of a high-pressure approach for school improvement, it must make intuitive sense in some circles that design educational policy at the present time. Otherwise we would not encounter

accountability policies with a heavy reliance on "sanctions as the fallback solution" (Mintrop & Trujillo, 2005). We, on the other hand, submit, as far as a study of this scale can do this, that it is the power of practically meaningful aspects of accountability combined with a supportive and open organizational climate and mild pressure that drive schools to grow, confirming findings from an earlier study (Mintrop, 2004). Even though successful schools in our case selection had a more positive attitude, educators in the nine schools overall had a dim view of the system's meaningfulness for their work. This is deplorable considering the positive improvement effects a more internalized approach could launch.

Accountability systems motivate educators to concentrate on student learning gains, and our success cases seem to exhibit such concentration. But the scope of their practices seems to revolve around a rather constricted notion of quality (one that excludes, for the most part, quality of teaching, for example), and they are encouraged to apply this notion in systems that reward strict alignment between content coverage and assessment. In order to foster the creation of better schools, rather than merely better aligned schools, designers need to widen the scope of quality and deepen the meaningfulness of the system for practice. We believe that these problems can be attenuated when systems become more open for a mix of quality indicators, perhaps some chosen by the state, some by the school. This openness is impossible as long as main drivers of school improvement are school rankings based on presumably ironclad performance indicators and the threat of sanctions.

As was mentioned before, almost all nine schools in our selection are faced with sanctions, mostly as a result of having failed to meet the more stringent federal AYP. Two of our three schools, identified as high-growth by this study, are slated to enter corrective action after having gone through the state's Underperforming Schools Program earlier. Sanctions make sense when people do not act responsibly, that is, when they willingly ignore justified expectations. This is clearly not the case in these two high-growth schools, nor in others that were less successful with their strenuous efforts. Our case studies show that subtle patterns of take-off, stalling, or coasting, countervailing the overall performance picture, may remain undetected by summary classifications of low and high performance. Decisions made on the basis of summary classifications—for example about the imposition of sanctions in the lockstep fashion of NCLB—may disregard these highly relevant patterns for school improvement. Our research suggests that accountability system designers ought to

raise the practical relevance of accountability systems for school improvement by introducing more fine-grained indicators of service quality and organizational health.

Limits of the Study

Studies such as this one are not designed to render generalizable findings, but they raise questions and direct our attention to patterns previously less seen. They help develop robust instruments and explore constellations with depth that then can be tested on a larger scale. Our findings are based on nine cases. As a result, identified patterns can depend on one or two cases or the specific groupings of cases we chose for comparison. This limits the stability of patterns uncovered. We attenuated this problem by emphasizing consistency across multiple measures and various group comparisons, but we obviously do not know how idiosyncratic or typical our nine schools are.

Our claims, restricted as they are, depend on an appropriate choice of measures of quality and organizational effectiveness. We tried to choose a number of conventional ones that have strong support in the scholarly literature and wide appeal to practitioners, but a potential "omitted variable bias" remains. For example, it is conceivable that instructional variables very proximate to content and its delivery, rather than the broader measures of time on task, cognitive complexity, student writing, engagement, and so on, would show a stronger association with higher standardized test scores. In this research, we were foremost interested in capturing some of these broader, more easily tangible characteristics of quality that transcend mere alignment; in a follow-up study we might add a number of more content-proximate measures.

As with all research, we wrestled with biases and skirted traps. For the sake of the immense effort that goes into accountability-induced school improvement, we hoped that an indicated performance gap that would take our lower performers 5 years or more to fill ought to result in better schools, not just better aligned ones. Consequently, we borrowed from research that assumes a process-product connection, such as the research on effective schools. On the other hand, we reminded ourselves that performance as indicated by the system could be mired in measurement error. Sure enough, we encountered many of the conditions that commonly weaken the validity of a performance index that is calculated on year-to-year averages. Some of our selected schools had their elementary school feeder

patterns changed, some deliberately started honors or magnet programs to attract higher performing students, and some lost funding or changed class size or turned over rapidly, making claims of consistent relationships between indicated performance and realities on the ground rather heroic. Thus, we needed to avoid the "causality trap" while at the same time searching for possible connections. In the end, we were unable to find solid connections in some areas (e.g., quality), but identified more consistent associations in others (e.g., accountability).

But in all of this, an element of uncertainty remains. We hope that we presented our findings in ways that allow the reader to cross-check our interpretations of the data. If we reinforced scholars' and practitioners' concern for consequential validity and practical relevance of accountability systems with our nine cases in one state, we have reached our goal. In this spirit, we now return to the initial practitioner paradox.

Revisiting the Paradox

Can we make a connection between system-indicated performance and school quality or is there little we can learn from accountability success cases? Broadly speaking, the nine schools are surprisingly similar in many dimensions of quality and in the instructional change strategies they employed, regardless of their success or failure in the accountability system, so encompassing claims of quality are probably not warranted. But there are some differences from which practitioners can learn, namely the way higher growth schools constructively engage with accountability.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14.

Achinstein, B. (2002). Conflict amid community: The micropolitics of teacher collaboration. *Teachers College Record, 104*, 421-455.

Ashton, P., & Webb, R. (1986). *Making a difference: Teachers' sense of efficacy and student achievement.* New York: Longman.

Blankstein, A. (2004). *Failure is not an option: Six principles that guide student achievement in high-performing schools*. Thousand Oaks, CA: Corwin Press.

Bomotti, S., Ginsberg, R., & Cobb, B. (2002, April). *Different teachers, different stakes? Determinants of attitudes toward high-stakes testing.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

California Department of Education. (2005). *California Basic Educational Data System (CBEDS)*. Sacramento, CA: Author. Retrieved 13 September 2006 from http://www.cde.ca.gov/ds/sd/cb/

California Department of Education. (2006). *API description: Overview of the Academic Performance Index (API)*. Sacramento, CA: Author. Retrieved 13 September 2006 from http://www.cde.ca.gov/ta/ac/ap/apidescription.asp

Carter, S. (2001). *No excuses: Lessons from 21 high-performing, high-poverty schools*. Washington, DC: Heritage Foundation.

Consortium on Chicago School Research. (2003a). *Survey of Chicago public school students, spring 2003, elementary student edition*. Chicago: Author.

Consortium on Chicago School Research. (2003b). *Survey of Chicago public school teachers, spring 2003, elementary teacher edition.* Chicago: Author.

Council of Chief State School Officers, Wisconsin Center for Education Research, & Learning Point Associates. (2003). *Surveys of enacted curriculum*. Washington, DC: Author.

Deal, T., & Peterson, K. (1991). *The principal's role in shaping school culture*. Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement, Programs for the Improvement of Practice.

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self determination in human behavior*. New York: Plenum.

EdSource. (2003). *California's lowest performing schools: who they are, the challenges they face, and how they're improving*. Mountain View, CA: Author.

Elmore, R. (2004). *School reform from the inside out: Policy, practice, and performance*. Cambridge, MA: Harvard Education Press.

Fitz-Gibbon, C., & Kochan, S. (2000). School effectiveness and education indicators. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 257-282). London: Falmer.

Fullan, M. (2003). *The moral imperative of school leadership*. Thousand Oaks, CA: Corwin Press.

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

Gray, J. (2001). Building for improvement and sustaining change in schools serving disadvantaged communities. In M. Maden (Ed.), *Success against the odds, five years on: Revisiting effective schools in disadvantaged areas* (pp. 1-39). London: RoutledgeFalmer.

Hallinger, P., & Heck, R. (1996). Re-assessing the principal's role in school effectiveness: A review of the empirical research, 1980-95. *Educational Administration Quarterly, 32*(1), 5-44.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Policy Analysis Archives, 8*(41). Retrieved 13 September 2006 from http://epaa.asu.edu/v8n41/

Hannaway, J., & Chaplin, D. (1994). *Breaking the cycle: Instructional efficacy and teachers of "at-risk" students*. Washington, DC: Urban Institute.

Haycock, K. (1999). *Dispelling the myth: High poverty schools exceeding expectations*. Washington, DC: EdTrust.

Hightower, A., Knapp, M., Marsh, J., & McLaughlin, M. (2002). The district role in instructional renewal: Making sense and taking action. In A. Hightower, M. Knapp, J. Marsh, & M. McLaughlin (Eds.), *School districts and instructional renewal* (pp. 193-201). New York: Teachers College Press.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal, 38,* 499-534.

Kane, T., & Staiger, D. (2002). *Volatility in school test scores: Implications for test-based accountability systems*. Washington, DC: The Brookings Institution.

Kim, J., & Sunderman, G. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3-13.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

LeCompte, M., & Dworkin, A. (1991). *Giving up on school: Student dropouts and teacher burnouts*. Newbury Park, CA: Corwin Press.

Linn, R. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives, 13*(33). Retrieved 13 September 2006 from http://epaa.asu.edu/epaa/v13n33/

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

Little, J. (1982). Norms of collegiality and experimentation: Workplace conditions of school success. *American Educational Research Journal, 19,* 325-340.

Louis, K., Febey, K., & Schroeder, R. (2005). State-mandated accountability in high schools: Teachers' interpretations of a new era. *Educational Evaluation and Policy Analysis, 27*, 177-204.

Louis, K., & Kruse, S. (1998). Creating community in reform: Images of organizational learning in inner-city schools. In K. Leithwood & K. S. Louis (Eds.), *Organizational learning in schools* (pp. 17-46). Lisse, Netherlands: Swets & Zeitlinger.

McBeath, J., & Mortimore, P. (Eds.). (2001). *Improving school effectiveness*. Buckingham, England: Open University Press.

McLaughlin, M., & Talbert, J. (1993). *Contexts that matter for teaching and learning: Strategic opportunities for meeting the nation's educational goals*. Stanford, CA: Center for Research on the Context of Secondary School Teaching.

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

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Mintrop, H. (2003). The limits of sanctions in low-performing schools: A study of Maryland and Kentucky schools on probation. *Education Policy Analysis Archives, 11*(3). Retrieved 13 September 2006 from http://epaa.asu.edu/epaa/v11n3.html

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

Mintrop, H., & MacLellan, A. (2002). School improvement plans in elementary and middle schools on probation. *Elementary School Journal, 102,* 275 -300.

Mintrop, H., & Trujillo, T. (2005). Corrective action in low performing schools: Lessons for NCLB implementation from first-generation accountability systems. *Education Policy Analysis Archives, 13*(48). Retrieved 13 September 2006 from http://epaa.asu.edu/epaa/v13n48/

Newmann, F., Bryk, A., & Nagaoka, S. (2001). *Authentic intellectual work and standardized tests: Conflict or coexistence?* Chicago: Consortium on Chicago School Research.

Newmann, F., Secada, W., & Wehlage, G. (1995). *A guide to authentic instruction and assessment: Vision, standards, and scoring*. Madison: University of Wisconsin, Wisconsin Center for Educational Research.

Newmann, F., Smith, B., Allensworth, E., & Bryk, A. (2001). Instructional program coherence: What it is and why it should guide school improvement policy. *Educational Evaluation and Policy Analysis, 23*, 297-321.

Newmann, F., & Wehlage, G. (1995). *Successful school restructuring: A report to the public and educators*. Madison: University of Wisconsin, Center on Organization and Restructuring of Schools.

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

Oakes, J. (2002). *Education inadequacy, inequality, and failed state policy: A synthesis of expert reports prepared for Williams v. State of California*. Los Angeles: University of California, Institute for Democracy, Education, & Access.

Odden, A., & Kelley, C. (1997). *Paying teachers for what they know and do*. Thousand Oaks, CA: Corwin Press.

Organisation for Economic Co-operation and Development. (2000). *PISA 2000 Technical report*. Paris: OECD Publications.

Ravitch, D. (1983). *The troubled crusade: American education, 1945-1980*. New York: Basic Books.

Reeves, D. (2000). *Accountability in action: A blueprint for learning organizations*. Denver, CO: Advanced Learning Press.

Rowan, B., Chiang, F., & Miller, R. (1997). Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of Education, 70*, 256-284.

Russell, M. (2002). *California's accountability system and the API, expert witness report for Eliezer Williams et al. v. State of California*. San Francisco, CA. Retrieved 13 September 2006 from http://www.decentschools.com/experts.php?sub=per

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

Scheerens, J. (1992). *Effective schooling: Research, theory and practice*. London: Cassell.

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness* (1st ed.). Oxford: Pergamon Press.

Sergiovanni, T. (1992). *Moral leadership: Getting to the heart of school improvement* (1st ed.). San Francisco: Jossey-Bass.

Skrla, L., & Scheurich, J. (2003). *Educational equity and accountability: Paradigms, policies, and politics*. New York: RoutledgeFalmer.

Springboard Schools. (2005). *Challenged schools, remarkable results: Three lessons from California's highest achieving high schools*. San Francisco, CA: Author.

SRI International, Policy Studies Associates, & Consortium for Policy Research in Education. (2003). *Evaluation of Title I accountability systems and school improvement efforts (TASSIE), 2002-03*. Menlo Park, CA: Author.

Stoll, L., & Fink, D. (1998). The cruising school: The unidentified ineffective school. In L. Stoll & K. Myers (Eds.), *No quick fixes: Perspectives on schools in difficulty* (189-206). London: Falmer Press.

Taylor, B. (2003). *School change classroom observation manual*. Minneapolis: University of Minnesota.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.

Teddlie, C., & Stringfield, S. (1993). *Schools do make a difference: Lessons learned from a ten-year study of school effects*. New York: Teachers College Press.

WestEd. (2005). *Schools moving up*. San Francisco, CA: Author.

Yin, R. (2003). *Case study research: Design and methods* (3rd ed.). Beverly Hills, CA: Sage Publications.

Appendix A

Teacher and Student Questionnaire Variables

Variable name	Definition
	Student Educational Experience
Academic Engagement	Students find classes interesting and challenging
Academic Press	Teachers have high expectations of students
Teacher Care	Teachers care for and listen to students
Peer Collaboration	Students like to work cooperatively
Safety	Students feel safe around the school campus
	Accountability
Goal Importance	Personal importance of accountability system and goals
External Validation	System supplies professional prestige
Authoritativeness	Teachers should comply with state or district mandates no matter what
Threat	Personal anxiety due to sanctions
Pressure	Accountability imposes pressure on school
Focus	System provides a focus for instruction
Diagnostics	System provides useful information to drive instruction
Validity	System is a valid gauge of teachers' performance
Fairness	System is a fair gauge of teachers' performance
Realism	System targets are realistic
Raised Expectations	Teachers expect and assign more challenging work
Goal Integrity	System goals and demands are balanced with teachers' values and student needs
Test Importance–Personal	Students feel high state test scores are personally important
Test Importance–Whole School	Students feel high state test scores important for the whole school
Sanction Awareness	Students are aware of consequences for low school performance
Test effort	Students push themselves when taking state tests
	Leadership
Urgency	Pressure for continuous improvement, reinforced by principal
Principal Support	Administration encourages and recognizes staff members for a well done job
Principal Control	Administration sets school priorities, makes and enforces plans
School Management	School is organized and functions well
Open Communication	Open discussions are encouraged and it is okay to disagree
Autonomy	Teachers' professional judgment and creativity are respected
Instructional Leadership	Administration sets high teaching standards and understands how children learn
Moral Leadership	Administration models how to put the needs of children first

Variable name	Definition
	Faculty Culture
Collegiality	Cooperative effort and support among staff
Pulling Together	Cooperative effort and support among staff driven by accountability demands
Norms of Performance	Teachers set and hold each other to high standards
Learning Orientation	Teachers continually learn and respect professional expertise
	Motivation
Involvement	Teachers' present level of involvement in improvement activities
Effort–1	Work hours increased due to school improvement efforts
Effort–2	Willingness to put in a great deal of effort beyond expectations
Hard Work	Teachers work beyond contractual hours
Commitment	Teachers have high commitment to stay at the school
Morale	Teachers believe school is on continuous improvement path
Satisfaction	Teachers feel satisfied with their work and the school
	Efficacy and Qualifications
Instructional Efficacy	Teachers can effectively reach even the most difficult students
Test-Related Efficacy	Teachers have knowledge and skills of how to do well on state tests
Colleagues' Skills	Colleagues are well prepared to meet performance expectations
Preparedness	Teachers feel well prepared for this year's teaching assignment
Years Teaching	Total years teachers have taught
Years at School	Total years teachers have taught at this school
Degree	Highest degree held by teachers
Full Certification	Teachers are fully certified to teach this year's assignment
	Change Strategies
Program Coherence	Continuity exists among programs
Strategic Orientation	School continually adjusts medium- or long-term improvement strategies
Money & Hopefulness	Low-performing schools funding has made me hopeful
Money & Impact	Low-performing schools funding has had some impact
Planning	School improvement plan provides a focus for school to carry out
Data Usage	Various sources of data are important for teachers' work
District Operational System	District provides consistent messages and aligns activities
District Instructional System	District provides useful instructional and curricular guidance
	Background
Familial Support	Parent or another adult helps and encourages students
Parent Support	Parents are involved in school activities
Possession of Cultural Goods	Students' families have newspapers, magazines, and a computer

Appendix B
Student Survey Scales

Student Educational Experience

Academic Engagement[a]	*Factor Loading*
Most of the topics we are studying are interesting and challenging.	.513
I usually look forward to most of my classes.	.572
I work hard to do my best in most of my classes.	.466
I am usually bored in most of my classes.	.472
Sometimes I get so interested in my work I don't want to stop.	.525
I often count the minutes until class ends.	.396
Most of my classes really make me think.	.480
Reliability (Cronbach alpha) = .69	

Academic Press[a]	*Factor Loading*
Most of my teachers:	
• expect me to do my best all of the time.	.573
• expect everyone to participate.	.538
• don't allow me to be lazy.	.486
• expect everyone to work hard.	.605
Reliability (Cronbach alpha) = .77	

Teacher Care[b]	*Factor Loading*
Students get along well with most teachers.	.482
Most teachers at this school care about students.	.600
Most of my teachers really listen to what I have to say.	.663
If I need extra help, I will receive it from my teachers.	.533
Most of my teachers treat me fairly.	.643
Reliability (Cronbach alpha) = .79	

Peer Collaboration[b]	*Factor Loading*
I like to work with other students.	.680
I learn most when I work with other students.	.652
I like to help other people do well in a group.	.567
It is helpful to put together everyone's ideas when working on a project.	.530
Reliability (Cronbach alpha) = .74	

Safety[a]	*Factor Loading*
How safe do you feel:	
• around the school?	.711
• in the hallways and bathrooms of the school?	.678
• in your classes?	.614
Reliability (Cronbach alpha) = .74	

Accountability

Sanction Awareness
Some students will transfer to other schools.
Teachers at our school will be transferred.
Our principal will be transferred.
The state or district will take control of our school.
Our school will be closed.
Scores calculated as the sum of the items.

Background

Familial Support[a]	*Factor Loading*
How often does a parent or another adult living with you:	
• help you with your homework?	.584
• check to see if you have done your homework?	.599
• tell you they are proud of you for doing well in school?	.624
• push you to take responsibility for the things you've done?	.640
• talk to you about working hard at school?	.695
• push you to go to college?	.577
Reliability (Cronbach alpha) = .79	

Possession of Cultural Goods
Does your family:
• get a newspaper at least four times a week?
• get any magazines regularly?
• have a computer at home that you use?
Scores calculated as the sum of the items.

[a]Adapted from Consortium on Chicago School Research (2003a).

[b]Adapted from Organisation for Economic Co-operation and Development (2000).

Appendix C
Teacher Survey Scales

Accountability

Goal Importance[a]	*Factor Loading*
It is very important for me personally that the school meet its state and federal performance targets.	.852
It really does not make much difference to me whether this school is (or may be) designated as an underperforming or program improvement school. (Values are reversed.)	.710
A high score on the state tests means a lot to me.	.820
It says nothing about me personally as a teacher whether the school raises the scores on the state tests or not. (Values are reversed.)	.691
Reliability (Cronbach alpha) = .76	

External Validation	*Factor Loading*
Meeting the expectations of the accountability system is a matter of professional pride for me.	.791
I work towards high test scores for our school because they enhance our standing in the district.	.887
It is important for me to meet our performance targets so that our school's reputation will not be damaged.	.883
Reliability (Cronbach alpha) = .81	

Authoritativeness	*Factor Loading*
Since California state authorities have decided to evaluate schools with the present accountability system, teachers ought to follow it.	.822
Teachers have little choice but to comply with state mandates.	.820
I implement state or district mandates even when they don't make sense to me personally.	.753
Reliability (Cronbach alpha) = .72	

Threat	*Factor Loading*
Sanctions:	
• make me more anxious for my career.	.903
• will have negative consequences for me personally.	.897
• put a lot of pressure on me personally.	.924
Reliability (Cronbach alpha) = .89	

Focus[a]	*Factor Loading*
State standards, tests, and performance targets:	
• provide a focus for my teaching.	.857
• tell us what is important for this school to accomplish.	.883
• have made us concentrate our energy on instruction and student learning.	.761
Reliability (Cronbach alpha) = .77	

Diagnostics[b]	*Factor Loading*
Results from state tests give teachers some useful feedback about how well they are teaching in each curricular area.	.840
Results from the state tests can provide valuable diagnostic information.	.893
The state tests provide little useful information for my instruction. (Values are reversed.)	.739
The state tests provide information that helps schools improve.	.875
State test results help identify students who need additional academic help.	.787
Reliability (Cronbach alpha) = .88	

Validity[a]	*Factor Loading*
The state assessments assess all of the things I find important for students to learn.	.788
A good teacher has nothing to fear from the state accountability system.	.775
The state assessments reflect just plain good teaching.	.843
Reliability (Cronbach alpha) = .72	

Fairness[a]	*Factor Loading*
For the most part, teachers are unfairly judged by the accountability system. (Values are reversed.)	.750
I resent being judged based on school-wide test scores and the performance of other teachers. (Values are reversed.)	.679
All schools in California have a fair chance to succeed within the state accountability system.	.643
The accountability system is stacked against schools located in poor communities. (Values are reversed.)	.719
Our students are not behind because of the teachers they have, but because of the conditions in which they have to grow up. (Values are reversed.)	.760
Reliability (Cronbach alpha) = .75	

Realism[a]	*Factor Loading*
The performance expectations of the state are for the most part unrealistic. (Values are reversed.)	.765
API targets are realistic goals for our school.	.797
AYP targets are realistic goals for our school.	.736
It is unrealistic to expect schools that serve poor neighborhoods to perform on the same level as schools in wealthy neighborhoods. (Values are reversed.)	.713
The state assessments are unrealistic because too many tasks are too hard for our students. (Values are reversed.)	.688
Reliability (Cronbach alpha) = .79	

Raised Expectations	*Factor Loading*
As a result of state standards, assessments, and accountability pressures:	
• I expect more from students.	.870
• I assign more challenging work.	.883
• I expect more from myself as a teacher.	.853
• I assign more complex cognitive tasks.	.831
Reliability (Cronbach alpha) = .88	

Goal Integrity
How important should these forces be?
District and state demands
Student needs
Teachers' values and goals
How important are these forces in reality at your school?
District and state demands
Student needs
Teachers' values and goals
Scores calculated based on differences between like items.

Leadership

Urgency	*Factor Loading*
The accountability system makes continuous improvement an urgent task for our school.	.770
Being held accountable by the state has made us aware of what we must accomplish at this school.	.698
The principal uses the pressures of accountability to move our school forward.	.781
The principal has encouraged teachers to see the accountability system as a tool for our school to improve.	.737
Reliability (Cronbach alpha) = .73	

Principal Support[a]	*Factor Loading*
The school administration's behavior toward the staff is supportive and encouraging.	.929
The principal usually consults with staff members before s/he makes decisions that affect teachers.	.904
Staff members are recognized for a job well done.	.905
Reliability (Cronbach alpha) = .90	

Principal Control[a]	*Factor Loading*
The principal sets priorities, makes plans, and sees that they are carried out.	.738
The principal puts pressure on teachers to get results.	.715
In this school, the principal tells us what the district and state expect of us, and we comply.	.856
Reliability (Cronbach alpha) = .64	

School Management	*Factor Loading*
This school is well managed.	.938
Overall this school functions well.	.920
Our administrators are good managers who know how to make our school run smoothly.	.932
This school is disorganized. (Values are reversed.)	.832
Reliability (Cronbach alpha) = .93	

Open Communication	*Factor Loading*
Open discussions about the meaningfulness of the state accountability system and related district policies are encouraged.	.823
Faculty gatherings provide a forum to discuss different perspectives on school improvement.	.880
It is okay to speak up when you disagree with the powers that be.	.862
Teachers are mainly encouraged rather than told to implement new programs or policies.	.792
Reliability (Cronbach alpha) = .86	

Autonomy	*Factor Loading*
Teachers' expertise in the classroom domain is respected here.	.842
In this school, I am encouraged to be creative in my classroom.	.860
In this school, I am given the space to exercise my professional judgment as to what is best for my students.	.851
Reliability (Cronbach alpha) = .81	

Moral Leadership
The administration at this school:
• places the needs of children ahead of personal and political interests.
• models the kind of school they want to create.
r = .75

Instructional Leadership[c]	*Factor Loading*
The administration at this school:	
• makes clear to the staff their expectations for meeting instructional goals.	.759
• sets high standards for teaching.	.860
• understands how children learn.	.831
• sets high standards for student learning.	.841
• broadly shares leadership responsibility with the faculty.	.684
• carefully tracks student academic progress.	.751
• monitors and evaluates the quality of teaching in a way that is meaningful for teachers.	.800
• allocates resources and other supports according to the school's goals and standards.	.746
Reliability (Cronbach alpha) = .91	

Faculty Culture

Collegiality[d]	*Factor Loading*
Most of my colleagues share my beliefs and values about what the central mission of the school should be.	.763
There is a great deal of cooperative effort among staff here.	.875
I can count on colleagues here when I feel down about my teaching or my students.	.805
In this school, the faculty discusses major decisions and sees to it that they are carried out.	.760
Reliability (Cronbach alpha) = .81	

Pulling Together	*Factor Loading*
At this school, when it comes to meeting the challenges of reaching our API or AYP targets, administrators and teachers are on the same side.	.799
Facing the pressures of school accountability has brought the faculty together; almost everyone is making a contribution.	.895
The pressures of meeting API or AYP targets have strengthened the hand of those at the school who are interested in good teaching.	.836
Reliability (Cronbach alpha) = .80	

Norms of Performance	*Factor Loading*
In your judgment, how many teachers at this school:	
• help maintain discipline in the entire school?	.730
• take responsibility for improving the school?	.875
• set high standards for themselves?	.886
• are eager to try new ideas?	.871
• feel responsible to help each other do their best?	.861
• feel responsible when students in this school fail?	.715
Reliability (Cronbach alpha) = .90	

Learning Orientation[d]	*Factor Loading*
My job provides me with continuing professional stimulation and growth.	.657
Teachers in this school continually learning and seeking new ideas.	.812
The staff seldom evaluates its programs and activities. (Values are reversed.)	.603
Teachers at this school respect those colleagues who are expert at their craft.	.804
The most expert teachers in their field are given leadership roles at this school.	.739
Reliability (Cronbach alpha) = .76	

Satisfaction
How often do you feel satisfied:
• with your work as a teacher?
• with your school overall?
r = .52

Efficacy and Qualifications

Instructional Efficacy	*Factor Loading*
I have found a way to get through to even my most difficult students.	.647
Sometimes I wonder if I would be more effective teaching a different age group. (Values are reversed.)	.646
In general, my classes are disciplined and well behaved.	.720
Students know that I expect hard work from them and they act accordingly.	.749
My challenge in this school, frankly, is to get through the day. (Values are reversed.)	.609
For the most part, my students are engaged in my lessons.	.730
Reliability (Cronbach alpha) = .75	

Test-Related Efficacy
I have the skills and knowledge needed for my students to meet the performance expectations of the state.
I know how to teach so that students will do well on the state tests.
r = .52

Colleagues' Skills[a]	*Factor Loading*
Most of my colleagues have the knowledge and skills needed for our school to meet the performance expectations of the state.	.827
The typical teacher at this school ranks near the top of the teaching profession in knowledge and skills.	.855
Many teachers in this school are insufficiently prepared to do their jobs well. (Values are reversed.)	.778
Reliability (Cronbach alpha) = .75	

Change Strategies

Program Coherence[c]	*Factor Loading*
Once we start a new program, we follow up to make sure it's working.	.784
We have so many different programs in this school that I can't keep track of them all. (Values are reversed.)	.777
Many special programs come and go at this school. (Values are reversed.)	.831
You can see real continuity from one program to another at this school.	.810
Reliability (Cronbach alpha) = .81	

Strategic Orientation
A medium or long-term strategy that keeps our school on a path of continuous improvement is clearly in place.
At this school, we adjust improvement strategies and programs to the varying needs of students or teachers.
r = .61

Data Usage	*Factor Loading*
Overall student performance on state or district tests.	.675
Student performance on state or district tests, disaggregated by class.	.674
Student performance on state or district tests, disaggregated by subgroup.	.697
Subtest or item-cluster scores on state or district tests.	.727
Item-by-item review of state or district test results.	.505
Student performance on school-level assessments (e.g., common writing prompts, math tasks, or reading assessments).	.572
Surveys of teachers, students, and/or parents.	.689
Information from classroom observations.	.538
Characteristics of students who are retained and/or drop out.	.640
Measures of school safety and discipline.	.671
Attendance rates.	.648
Student mobility rates.	.631
Reliability (Cronbach alpha) = .87	

District Operational System[e]	*Factor Loading*
Our district:	
• monitors our progress on goals established in our school plans.	.739
• sends consistent messages regarding our school goals and improvement strategies.	.849
• provides adequate assistance for our school's improvement.	.914
• provides useful feedback on our school improvement efforts.	.898
• proposes improvement activities that are in line with our goals.	.905
• has standardized instructional approaches for our school.	.576
Reliability (Cronbach alpha) = .90	

District Instructional System	*Factor Loading*
Our district provides:	
• useful reports of student achievement data.	.687
• clear guidance on what curriculum we should teach.	.786
• clear guidance on how we should deliver our instruction.	.788
• effective professional development that helps our school reach its goals.	.748
Reliability (Cronbach alpha) = .75	

Background

Parental Support[c]	*Factor Loading*
At this school, how many of your students' parents:	
• attend parent-teacher conferences when you request them?	.713
• return your phone calls promptly?	.770
• attend a sports event on campus?	.505
• attend a student performance on campus?	.670
• attend Back-to-School Night?	.696
• support your teaching efforts?	.787
• do their best to help their children learn?	.748
Reliability (Cronbach alpha) = .83	

[a]Adapted from Mintrop (2004).

[b]Adapted from Bomotti, Ginsberg, and Cobb (2002).

[c]Adapted from Consortium on Chicago School Research (2003b).

[d]Adapted from McLaughlin and Talbert (1993).

[e]Adapted from SRI International, Policy Studies Associates, and Consortium for Policy Research in Education (2003).

Appendix D
Classroom Observation Measures

Name	Definition
	Percentage of snapshots in which:
Non-instructional time	Classroom activity was not related to student learning
Time on task	At least three fourths of students were on-task
Student engagement	Students appeared highly engaged in the lesson
Positive teacher tone	Teacher communicated with students using a positive, engaging tone (e.g., warm, task-oriented, inspired)
Proactive instruction	Teacher employed active instructional techniques (e.g., modeling, coaching, recitation, discussion, assessment)
Cognitive complexity	Students engaged in cognitively demanding activities (e.g., demonstrate/explain; analyze/investigate; evaluate; generate/create)