Educational Evaluation and Policy Analysis
December 2007, Vol. 29, No. 4, pp. 319–352
DOI: 10.3102/0162373707309219

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

Heinrich Mintrop
University of California, Berkeley
Tina Trujillo
University of California, Los Angeles

In search for the practical relevance of accountability systems for school improvement, the authors ask whether practitioners traveling between the worlds of system-designated high- and low-performing schools would detect tangible differences in educational quality and organizational effectiveness. In comparing nine exceptionally high and low performing California middle schools, the authors conclude that if such travelers expected to encounter visible signs of an overall higher quality of students' educational experience at the high-performing schools, they would be disappointed. Rather, they would have to settle on a narrower definition of quality that is more proximate to the effective acquisition of standards-aligned and test-relevant knowledge. High-growth schools tended to generate internal commitment for accountability and consider it an impetus for raising standards.

Keywords: *school quality, state accountability systems, school improvement, school effectiveness*

THIS article grows out of two motifs that have surfaced repeatedly in conversations with school practitioners and students in the Principal Leadership Institute (PLI) at the University of California, Berkeley, 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 schools, PLI students would list their schools' demographics, the likable features of school life, perhaps the relationships between principals and teachers, and end up at some major challenges. In 2005, after 6 years of state accountability, PLI students most often state that their schools are 1-1 schools, 3-5 schools, schools 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 on the basis of this indicator (see below for more details). Within 5 short years, those numbers seemingly have become signals of schools' quality and character, an increasingly powerful shorthand and social fact in the lives of school practitioners. Is this justified?

We would like to acknowledge the generous support of the Center for Research on Evaluation, Standards and Student Testing at the University of California, Los Angeles, for this research and the valuable and helpful comments from Bruce Fuller, Joan Herman, Betty Malen, the *EEPA* editors, and the three anonymous reviewers. Yeow Meng Thum helped us with case selection procedures. Thanks also to Rei Chan, Kris Kim, Corey Triplett, Alejandra Velasco, and Tanner LeBaron Wallace for their contributions to this work.

Increasingly, administrators and teachers at 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. Yet we frequently encounter the assertion from practitioners at 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 Paradox of Accountability Systems

The two motifs are paradoxical: Either the system-based performance categories stand for some broader characteristics of school quality, or they are not connected to tangible educational and organizational conditions from which educators can learn. Our research explores this paradox. It takes from conversations with practitioners that a connection cannot be taken for granted, and yet such a connection is intuitively made. This "practitioner paradox" relates to a deeper "system paradox" that is rooted in the disconnect between assessment and accountability functions: Accountability systems assess the quality of children's learning (Carlson, 2006), but it is to a large degree adults and their educative actions that are held to account. Performance information (e.g., test scores, performance indexes) are part and parcel of a strategy for improving educational quality, yet the systems' assessments are generally not designed to capture the quality of educational processes. Accountability systems virtually compel adults to draw inferences from student learning for adult action (Baker & Linn, 2004), hence to apply to them purposes for which they were not explicitly designed. Yet without such inferences, the accountability function would be rather pointless.

But are such inferences for adult action justified? Although statisticians and technical design experts show us that an accountability system may be solid in its student assessment function when standardized tests and performance indexes are reliable and valid expressions of student learning, practical urgency demands that we go beyond this narrow validation and explore to what degree such systems justify commonly made inferences for adult practice, pertaining to educational processes and school quality. Thus, we need to find out not only if accountability systems are valid but also if they have *practical relevance* for school improvement, simply because this is how they are used in practical life. An accountability system attains practical relevance when performance indicators and practitioner experience of quality and effectiveness speak to the same reality, that is, when school people draw justifiable inferences from their schools' performance status for the quality of their own schools and their own actions. In reporting our findings, we adopt the lens of some imagined practitioners who, on the basis of information from the state accountability system, travel to a number of schools of varying performance statuses to learn what to do. We accompany these travelers in our role as researchers and assist with robust instruments and the benefit of systematic inquiry.

The California Accountability System

In 1999, California followed in the footsteps of other states with the installation of its own outcome-based accountability system (Goe, 2004; Mintrop & Trujillo, 2005). The Public School Accountability Act, composed of a set of rewards, sanctions, and supports, established a system for holding schools accountable for reaching achievement goals. At the core of the act was the API, a numeric index from 200 to 1,000 that measures both school performance and growth (California Department of Education, 2006). Each school's annual API is based on a formula that calculates the weighted average of students' scores across content areas on criterion- and norm-referenced tests. Annual target API scores, calculated as 5% of the difference between a school's API and the statewide target, are intended to gradually move all schools to the state's target of 800 API points. The state also ranks schools in deciles from 1 to 10, both across the state (comparing each school to other schools statewide) and to similar schools (comparing each school to 100 other schools with similar demographic characteristics). For instance, a 2-10 school would rank among the bottom 20% statewide but in the top when compared with schools with a similar demographic makeup. Schools are identified as low performing and participate in the state's low-performing-schools program on the basis of their state ranks, not their similar-schools ranks (Goe, 2004).

The reliability and validity of state accountability systems have been discussed on a number of grounds (Baker & Linn, 2004). Here is not the place to review the discussion among the designers and critics of the California system. Some have shown that the reliability of the API is fairly robust (Hill, 2001; Rogosa, 2003; Rogosa & Haertel, 2003), warranting its use for indicating school performance. Others have raised concerns that should be taken into account in an interpretation of API scores. The system's consistency has been questioned as the types of tests used and the weights attached to them have changed, away from norm-referenced tests, such as the Stanford 9, to the California Standards Tests, which are criterion referenced, subject matter based, and better aligned with state standards (Russell, 2002). Moreover, the fairness and validity of testing English-language learners, abundant in the state's schools, in a language that they poorly understand has also been questioned (Abedi, 2004).

Linn, Baker, and Betebenner (2002) discussed the consequences of plotting school improvement trajectories through average student performance from year to year, thereby falsely assuming population stability over time. California, not unlike many other states, uses such a procedure. Measurement errors have been found to make reliable classifications for schools with typically small API movements and small populations uncertain (Kane & Staiger, 2002). Performance bands, such as API decile ranks, have been found to be fairly reliable (Hill, 2001). But API state rank needs to be read with caution, because it indicates a school's standing relative to other schools for any given year. Thus, a school can improve from year to year on the API without moving up in its API decile rank if other schools improved relatively more vigorously (Goe, 2004). Similar-schools ranks are less reliable and more volatile from year to year (Hill, 2001). Some of the demographic data on which similar-schools ranks are based are self-reported, which may add an element of uncertainty to the calculation.

We tried to attenuate some of these problems in our study by using for our analysis API scores from demographically similar middle schools that are fairly large in size and contrast widely across the state's performance spectrum in absolute API and growth on the index over time. We ignored the similar-schools ranks and report absolute state ranks as additional illustrative information that may bolster a school's claim of high performance. Most of our findings on API growth relate to a time period when the state standards tests had become the bedrock of the index. But uncertainties remain. The state has elected to use the API as an authoritative measure of school performance and has thus created a new reality for school practitioners, which they fill with their own presumptions and sense making. Our study aimed to explore this reality, not the system's validity per se. Thus, to enter into the realm of adults' presumptions and inferences about their own actions within the system, we gave the system the benefit of the doubt despite misgivings, not unlike many school people for whom the API is a "given" (Malen & Muncey, 2000) that structures their challenges.

Practical Relevance

From the perspective of the system's practical relevance for school improvement, we assumed 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 can be exploited by reformers to good ends. Characteristically, accountability systems sort schools according to high or low performance on the basis of quantitative performance indicators. Schools that have shown high growth are touted as exemplars (Carter, 2001; Haycock, 1999; Reeves, 2000) and are 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 confer public value on schools and have entered educators' minds, if not hearts. Because reaching target scores is of paramount importance to schools' organizational survival and standing in their districts and communities, the system is attaining increasing evaluative and self-evaluative power. In the practical world of high-stakes school improvement, the system's power becomes such that inferences from student performance for adult action are made regardless of the system's technical design and esoteric statistical properties.

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

Such substantiation, as far as it concerns the business of school improvement, seems to hinge on two fundamental claims:

- The state performance scores justify inferences on school quality.
- High growth on state indicators over time is not a chance occurrence but a reflection of superior school improvement efforts that can be emulated by lower performing schools.

These two claims are of utmost relevance for practitioners' efforts, particularly in schools that traditionally occupy the bottom rungs of the social and educational status hierarchy and find themselves in dire need for ideas on how to improve.

As school improvement researchers, we deliberately took on the accountability system as received by lay practitioners, not as intended by statisticians or technical system designers. In keeping with the parlance of the state, we looked at "API score" and "API growth" over time. In our search for practical relevance, we asked whether practitioners traveling between the worlds of API-designated high- and low-performing schools would detect tangible differences by observing concrete behaviors, looking at students' work, or inquiring about teachers', administrators', or students' perceptions. Would they see real differences in school quality? Would they find schools that are truly more effective?

Practical relevance can be defined in a variety of ways, because varied constellations of factors, poorly understood by the research literature, may produce high performance (Hanushek, 1994; Scheerens & Bosker, 1997). Our definition for this study blends common sense with the findings and conventions from effective schools research (see more details in the next two sections). But other models could be imagined and tested. Our model hinges on five conditions. Accountability systems have high practical relevance when designated high-performing schools (a) have higher achievement, as measured by the performance indicators; (b) face similar educational challenges, as roughly indicated by similar proportions of students from poor, ethnic, and linguistic minority backgrounds; (c) provide a higher quality educational experience for students; (d) function more effectively; and (e) engage with accountability mechanisms more productively. Often, we assume that high-performing schools, indicated as such by their accountability system, satisfy these conditions. At times, when our initially mentioned practitioners speak of their schools with shame or pride as 1-1 versus 3-3 schools or 643 versus 517, 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 demographically similar schools and groups of schools with sufficiently distinct performance scores and compare "high" and "low" performers on the quality of students' educational experience, organizational effectiveness, and engagement with the accountability system. This exploration of schools is guided by three research questions:

1. Is absolute performance level, as measured by the state indicator, matched by other quality criteria 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 been shown to play a prominent role in effective school improvement?
3. Are the specific mechanisms of the accountability system instrumental for school improvement?

Educational Quality

The relevance of accountability systems for school improvement increases when the systems' success cases (in this case, higher absolute API scores) also rate highly on other important quality indicators not measured by the systems. For this study, they are how students experienced their

schools, what kind of teaching they encountered, and what kind of work they produced in their classrooms.

Thus our data consist of students' perceptions of their schools, lesson observations, and student writing samples. We asked about students' perceptions of academic engagement, academic press, teacher care, peer collaboration, and safety. These are conventional variables that have been shown to be of high 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 were defined as ones at which students felt engaged and challenged but at the same time safe, cared for, and collaborative.

For lesson observations and student work samples, we concentrated 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 the 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 tried to ascertain degree of complexity in constructing arguments and interpreting phenomena (Newmann, Smith, Allensworth, & Bryk, 2001; Newmann & Wehlage, 1995). Last, we used student suspension rates in conjunction with perceptions of safety as a proxy for student discipline. Thus, with our quality indicators, we explored how schools were doing with regard to basic order and basic skills but also with regard to more advanced and complex learning and attitudes.

Organizational Effectiveness

Following the research on school effectiveness and school improvement, we hypothesized that if an accountability system worked properly, we would find organizational characteristics in the system's success cases (i.e., schools with high growth in API scores over time) that are commonly associated with school effectiveness. Indeed, the underlying logic of the system's accountability function (e.g., a focus on instruction, clear goals, frequent assessments) stands in the tradition 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), among which we selected a few salient but fairly conventional ones that appeared most relevant to extant conditions of schools under accountability.[1]

Time and again, 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 a 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*; others emphasize his or her role as a strong initiator and an enforcer of rules, particularly in often chaotic, urban, high-poverty environments (Teddlie & Reynolds, 2000). Although there seems to be a tendency for some leaders, coming under strong accountability pressure, to increase *control* and reinforce a 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 roles.

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 workforce with high *involvement* and the willingness to exert great *effort*. But challenge, and concomitant stress, need 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 (Ingersoll, 2001; LeCompte & Dworkin, 1991; 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 collective sense of *colleagues' skills* and *test-related efficacy* may play a role as well.

Recent literature on school improvement has pointed to the salience of a school's instructional management for a school's chance to improve. We selected *instructional program coherence* (Newmann et al., 2001), *data use* for evidence-based decision making (Blankstein, 2004), school improvement *planning* (Mintrop & MacLellan, 2002), and a *strategic orientation* toward change (McBeath & Mortimore, 2001) as aspects that capture the presumed rationality of the accountability-driven change model. On the basis of 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

Naturally, the practical relevance of accountability systems is tied to how schools pick up on the systems' 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 (Goe, 2006; O'Day, 2002) emanating from the systems 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 systems. 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 based our exploration of schools' responses to accountability on the following model (Mintrop, 2004). Schools attach varying degrees of *goal importance* to the demands of an accountability system. Importance could be more externally or internally motivated (Deci & Ryan, 1985). In an external nexus, teachers could calculate extrinsic rewards, such as the enhancement of professional prestige or the aversion of disadvantages, that is, they would act primarily out of a sense of *external validation*. They also may 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 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*.

Contrasting with these primarily external motives to heed accountability demands could be more internalized motives. The usefulness of a 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. The usefulness, rightfulness, and *realism* of targeted goals are the tripod on which the effectiveness and steering capacity of a performance indicator rests (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 systems 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 are described for heuristic purposes. The primary mechanism of accountability power in a given school may be pressure and threat. Principals may seize on these pressures, reinforce urgency or even fear, increase control, and tighten up the organization and instructional program. In contrast, schools may accept accountability systems as meaningful

guides; 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).[3]

Method and Data

Our study followed in the footsteps of an earlier study that explored the connection between district-level administrators' judgments about school quality and actual API scores (Baker, Goldschmidt, Martinez, & Swigert, 2002). That study found a correlation between such judgments and the schools' API scores, but it is not clear if district administrators knew the schools' API scores when they made these judgments. Our way of studying "practical relevance" was to look at a relatively small number of schools in depth that differed in absolute performance and growth over time as measured by the prime state indicator, the API, and then make blind judgments on the basis of a battery of data from the schools. 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 on the index over time but were as similar as possible with respect to social background. The study used mixed methods and drew from multiple data sources: statistical analysis of teacher and student survey data, quantitative and qualitative analyses 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' responses to a 340-item questionnaire, 4,148 students' responses to a 50-item questionnaire, 270 observed lesson segments in English language arts, 390 pieces of student work, and 157 interviews. The study used a structured, multiple-cases design that allowed for quantitative and qualitative cross-case comparisons (Miles & Huberman, 1994, chaps. 7–8; Yin, 2003). It used descriptive statistics and significance tests and applied 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 2004–2005 school year.

We 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 other researchers in the field; some were specifically developed for this study.[4]

We selected our cases on the following criteria:

- demographic similarity;
- below-state-average baseline (1999) performance (first to fourth decile);
- similar starting API score in 1999 at the inception of the system;
- significant difference in absolute performance levels at time of data collection; and
- contrasting high or low growth on the API over a period of 4 years.

Initially, we identified schools with exceptionally high and low growth on the API by predicting annual API achievement on the basis of school background characteristics and by 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, controlling for school characteristics.[5]

Schools we selected from the high and low ends of the performance spectrum had similar baseline API scores 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 (FRPL) participation; 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 below 500 and 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 were

TABLE 1
Academic Performance Scores of the Nine Selected Cases

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

Note. API = Academic Performance Index.
a. $M = 628$, $SD = 41.5$. The mean was calculated as the unweighted average of the nine schools' API scores and was slightly biased. The unbiased mean of the high and low groups was 624. Significant differences between the two performance groups were tested using the Wilcoxon-Mann-Whitney test ($z = –2.47$, $p = .0135$).

classified as low performing and five as high performing. Recruitment of schools was a challenging process and took longer than expected. In the end, we settled for nine schools. Because of intensifying accountability pressures, a large number of low-API schools declined to participate. 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 were biased toward 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 differed 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 differed by more than half a standard deviation from the nine-school mean in absolute API performance. Overall, the mean 2005 API score was 660 for the high group and 587 for the low group, a 73-point difference that was statistically significant at the .01 level. In comparison, the mean 1999 API score for the state was 631. By 2005, the mean state API score had grown 78 points to 709, though these figures were influenced by higher mean elementary API scores and much lower secondary ones. Movement in state ranks corroborates these group differences. All nine schools started in 1999 in the lowest, second lowest, or third lowest API decile. Six years later, the four schools classified as low had declined, whereas the five schools classified as high had moved up at least one decile; one school moved up three deciles, and one school remained in the same rank.

Although 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 showed statistically significant differences across groups, though they differed within groups. Three of the four schools in the low group appeared to be economically more challenged, as indicated by higher FRPL participation, whereas schools in the high group had higher proportions of English language learners. Two 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 finer grain size, and to increase confidence in our school matchup, we went beyond state-reported data and inquired about students' and teachers' perceptions of family background and support for education. Three of the four scales showed statistically significant differences across groups (see Table 3), but schools classified as low were more challenged only in the area of parental support (because of the low ratings of one school), whereas the higher performing schools were more challenged in the areas of language and the possession of cultural

TABLE 2
Demographic Characteristics of the Nine Selected Cases, 2004–2005

	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 or reduced-price lunch (%)	97	59	100	100	77	85	83	69	78
Parent education[a]	1.81	2.13	2.09	2.25	1.81	2.02	2.09	2.18	2.03

Source. California Department of Education.
Note. All means were statistically insignificant between the high and low groups using the Wilcoxon-Mann-Whitney test.
a. 1 = *not a high school graduate*, 5 = *graduate school*. Parent education is subject to the inaccuracies of self-reported data.

TABLE 3
Teacher and Student Perceptions of Family Background

	Low				High				
Mean Response	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** (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** (1 = *never*, 4 = *always*)	3.0	2.7	2.7	2.9	3.3	3.2	3.2	3.0	3.4

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

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 seemed fairly well matched demographically. In conclusion, our nine-school sample was a suitable case selection for our intended analyses. It showed substantial differences in student achievement as expressed by API performance but was sufficiently similar demographically to permit meaningful comparisons.

Analytical Procedures

We analyzed our quantitative and qualitative data with three procedures. First, we conducted blind ratings of schools with the help of case-ordered descriptive metamatrices. Second, we conducted statistical significance tests across various school groupings. Third, we investigated configurations unique to individual schools using both quantitative and qualitative data. Our research team consisted of seven people. To avoid bias, one member of the team only rated anonymized writing samples (the second rater had visited schools), one team member prepared decision matrices for blind ratings but was not involved in the actual ratings, and two team members analyzed quantitative data with concealed case IDs and performance statuses prior to having access to the qualitative data.

Descriptive metamatrices

These kinds of matrices (Miles & Huberman, 1994, pp. 190–192) display records for each individual case for decision making. In constructing the matrices, we initially concealed school IDs and focused only on the quantitative data. We then constructed, on the basis of measures from student questionnaires, teacher questionnaires, classroom observations, and writing sample ratings, a matrix that indicated whether a given measure of school quality or effectiveness fell above (indicated with a plus sign) or below (indicated with a minus sign) the mean. Asterisks were used to denote borderline cases, thereby alerting us to possible classification uncertainties. We represented these data in a matrix we called a school profile. This profile was fairly unbiased in that students' perceptions and teachers' responses 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 and effectiveness. This involved judgments not unlike ones that would have been made by our imagined traveling practitioners. Two members of the research team, the authors of this article, who had visited the schools during data collection but did not know which matrix belonged to which school, studied each school profile and then judged whether a school was a likely high or low performer or whether the case was undecided. The two raters had together about 18 years of practical experience in schools prior to becoming researchers. As a decision rule, we elected that if we could identify at least half of the schools correctly (in the high group, three out of five) and the rest at least as undecided without interrater disagreement, we had found sufficient justification for inferring school quality or effectiveness from API performance status or API growth, respectively.

Statistical significance tests

The purpose of statistical significance tests in our study was not to arrive at generalizable findings, nor to calculate exact estimates, but to increase our confidence in the subjective rater judgments by interrogating the data with yet another method. We conducted various significance tests to assess the differences between our low and high groups in terms of teacher survey responses, student survey responses, classroom observations, and student writing samples. Where appropriate, we weighted our data and accounted for the dependence of nested observations by treating the school as the cluster sampling unit. We also checked for the dependence of our measures by examining correlations, most of which were low, moderately low, and insignificant. When data were weighted, we could not conduct traditional independent-samples t tests to check for differences in means between performance groups. In their place, we performed weighted survey regression analyses (the standard procedures to use in place of t tests when using probability weights) in which we dummy coded a performance group predictor variable and accounted for our weights. We interpreted the t statistic and its p value in each test.

Individual schools

In a third step, we unveiled case IDs for the research team, analyzed interview data, and composed narratives that drew from both qualitative and quantitative data. The full analysis of these data must be left to another article.

Findings: Testing the Model of Practical Relevance

We undertake our analysis in two steps. First, we look at the relationship between *absolute* API performance and other educational quality measures that capture various aspects of students' educational experience. Second, we look at the relationship between API *growth* differentials and organizational effectiveness and accountability scales, in keeping with the assumptions of our model that higher growth in API over time should be reflected in a better functioning organization and a more productive response to accountability.

The Quality of Students' Educational Experience

Is absolute performance level as measured by the state indicator matched by other quality

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: +)			–	–	+	+	+		+
Noninstructional 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. ↑ = possibly high, ↓ = possibly low, 0 = undecided. API = Academic Performance Index.
a. Interrater disagreement.

criteria that circumscribe students' educational experience? As a reminder, our high group differed from our low group by a mean of 73 API points, and the difference between our top and bottom schools was 110 API points. These differences are not trivial given that it took our low-growth schools 6 years to make gains of 70 to 100 points.

Blind ratings

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 1 standard deviation of the nine-school mean and a plus or minus sign to means that fell more than 1 full standard deviation above or below the mean. 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 ordinal data (e.g., suspension rate, noninstructional time, student engagement) we divided the range of scores into four equal intervals and assigned plus or minus signs to scores in the top and bottom intervals and zeros for scores in the middle. For all ratings, we assigned asterisks to denote borderline cases to flag possible classification uncertainties.

Table 4 displays our decision-making matrix and ratings with interrater agreements, matched with the (previously concealed) absolute API scores at the time of data collection. As can be seen, we could not classify correctly and reliably a sufficient number of schools in the appropriate performance status group. Negative ratings for the quality of students' educational experience were more frequent in the low group, but one school (D) was rated as high by both raters. Schools in the high-performance group were not consistently rated better in terms of students' perceptions, classroom teaching, and the quality of student writing, standing together, with the exception of the top-API school (B).

Significance tests

We conducted significance tests for all of our student perception scales. As explained previously, we conducted weighted survey regression analyses to assess the differences between our performance groups. None of the tests turned up statistically significant differences of means.

TABLE 5
Measures of Quality of Students' Educational Experience

	Low				High				
	F	D	I	C	H	G	A	E	B
Student perceptions (*M*)									
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 (%)	44	32	59	48	9	21	21	25	13
Enacted curriculum (% of lesson snapshots)									
Noninstructional 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

Table 5 shows that as far as students' 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 experiences provided to them at 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 at the high-API schools.

As to classroom observation measures, Table 5 suggests that no consistent patterns were obtained for the quality of lessons across our high- and low-performance groups.[6] Indeed, these measures were all statistically insignificant using the Wilcoxon-Mann Whitney test (the nonparametric version of an independent-samples *t* test). Thus, blind ratings, significance tests, and descriptive analyses did not render 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 noninstructional 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

Although consistency in educational quality between our two API status groups was difficult to establish, the top-API school (B) stood out with positive marks on 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 was not the only one that was so rated. Indeed, low-API School D also was rated as high by the blind raters, primarily because of a higher frequency of lessons (strictly speaking, lesson snapshots) in which the teacher's tone was positive, the teaching was proactive, and learning activities went beyond mere recall. Thus, instructional quality, at least in the English language arts classes, was remarkably similar in the two schools, irrespective of a formidable

110-point difference in API score at the time of data collection.

Two schools in the low-API group, I and C, stood out with exceptionally high suspension rates and somewhat lower student safety ratings. This may hint at a higher disciplinary burden compared with schools in the high-API group (particularly Schools G and A). Whether this condition was due to the schools' own doing or social context is not entirely clear, but extremely high FRPL participation rates at these two schools (see Table 2) pointed in the direction of social-context differences that created challenges less encountered at the higher API schools. School F seemed to exhibit a similar pattern, albeit not as pronounced, as to student safety. Qualitative data confirmed a much higher concern for, and effort expended on, safety and order at three of the four low-API schools (I, C, and F). On the other hand, two high-API schools, A and G, stood out as very safe and orderly schools, though they were also highly affected 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 were not closely matched. Our traveling practitioners would indeed have a hard time distinguishing system-designated high-performing from low-performing schools by observing English teachers and by talking to students. They might gain a slightly better idea by examining student writing. But without knowing actual test results, they may likely lump schools from different performance groups together.

Organizational Effectiveness and Response to Accountability

Is high or low API growth as measured by the prime state indicator matched by organizationally more effective adult interactions and more productive responses to the accountability system? Exploring the practical relevance of accountability systems in the area of organizational effectiveness required 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 effectiveness of adult interaction. Hence, we selected into our high- and low-performance groups schools that not only had substantially different absolute API levels but that also arrived at these levels because of either exceptionally high or low API growth over time.[7]

Blind ratings

Again, we constructed a school profile matrix, this time on the basis of measures from teacher questionnaire data, assigning zeros (suppressed) to school means that fell within 0.1 point of the nine-school mean, one plus sign or one minus sign to means that fell less than 1 standard deviation above or below the mean, two plus signs or two minus signs to means that fell more than 1 full standard deviation above or below the mean, and asterisks to denote borderline cases.

Table 6 shows the blind ratings, drawn from Table 7, juxtaposed with (previously concealed) API growth differences. Despite the many more measures to be considered, raters disagreed on only two schools. We first compared schools across our original high-low distinctions. Clearly, the raters were unable to classify the schools correctly. More schools in the high groups were classified as less effective than more effective, and two schools in the low groups were rated as more effective.

Significance tests

To double-check our subjective ratings, we conducted significance tests for all of our teacher perception scales. Because these data were weighted to adjust for differences in school size, we conducted weighted survey regression analyses in which we treated the school as the cluster sampling unit to account for the dependence of our nested observations (see above). When we compared the five schools originally classified as high with the four classified as low, none of the school effectiveness and accountability measures listed in

TABLE 6
Academic Performance Index (API) Growth Ratings on the Basis of 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
2003–2005 API difference	36	–4	65	56	47	–4	37	36	78
Blind summary ratings	0/↑	↓	0	↑	↓	0↓	0	↓/0	↑

Note. ↑ = possibly high, ↓ = possibly low, 0 = undecided.
a. Interrater disagreement.

the school profile (Table 7) were significantly different, with one exception.[8]

Searching for a better fit

On the other hand, even a cursory look at the school profile matrix tells us that some schools were more effective than others according to our measures. Schools B and C stood out as the two cases that were rated unambiguously by the two blind raters as high growth. Examining a number of performance patterns, we found that these two schools grew rapidly on the API in the last 2 years. Also during these last 2 years, some previously classified high-growth schools had declined, whereas some low-growth schools had soared. Apparently, things had 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, was two lifetimes, given leadership changes, teacher turnover, student mobility (particularly in middle schools), and the fluid social composition of California immigrant communities. As a result, students' and teachers' perceptions about their schools in 1 year are rarely good for much longer.

When tested for change over merely 2 years, some organizational characteristics showed systematic relationships to API growth. Here too, we conducted weighted survey regression analyses to assess the differences between our new performance groups. Table 8 displays the results of these tests. The independent variable was dichotomous and consisted of two performance groups: three schools (B, C, and I) that posted the highest API score differences from 2003 to 2005 and five schools that posted substantially lower API score differences (A, D, E, F, and G), leaving out School H to provide for sufficient distance between high- and low-API-growth schools (see Table 6). The dependent variables were continuous perception scales. Only scales that showed statistical significance are displayed. 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 small number of cases, results can sway depending on what schools one decides to group together. The groups compared in Table 8 differed 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. However, the results remained fairly similar when comparing more extremely composed groups, for example, the three top-growth and two bottom-growth schools, suggesting a more stable pattern.

Although the means for many of the scales were fairly close together and hovered around scale midpoints, the significant measures, together, spoke to a conspicuous pattern: Rather than generic characteristics of effectiveness (e.g., 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 with similarly situated

TABLE 7

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	+	–		++			+	– –	++
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. (+) = means that fell less than one standard deviation above the nine-school mean; (–) = means that fell less than one standard deviation below the nine-school mean; (++) = means that fell more than one full standard deviation above the nine-school mean; (– –) = means that fell more than one full standard deviation below the nine-school mean; (*) = denotes borderline cases.

TABLE 8
Organizational Characteristics of Higher and Lower API Growth Schools, 2003–2005 (survey linear regression)

		Estimated Mean	
	Range	Lower Five	Higher Three
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

Note. API = Academic Performance Index.
*$p < .05$. **$p < .01$. ***$p < .001$.

lower API growth schools, schools that grew strongly over the last 2 years attached greater importance to accountability goals. They were more concerned about their external reputations and were more willing to accept the normative authority of the state to direct them. But by the same token, they regarded the system as more meaningful, albeit somewhat modestly: They saw it as more useful and had fewer doubts about its rightfulness, the latter indicated by scale means for "validity" and "focus" barely in the affirmative range. Accountability demands led them to raise expectations for themselves and their students. System demands, student needs, and their own values agreed more strongly. They pulled together as a faculty around accountability demands but at the same time had open discussions about accountability and chances to disagree. Teachers perceived their schools as better organized with regard to instructional program and change strategies. They did 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 used the accountability system to move an improvement agenda forward, some skepticism notwithstanding.

Individual schools

A look at individual schools (see Table 7) clarifies, refines, but also questions the pattern derived from group comparisons. Two of the top three high-API-growth schools (B and C) conformed most closely to the hypothesized model of organizational effectiveness and constructive engagement with accountability. Indications of meaningfulness and internalization of accountability were higher, although even in these schools, the strength of agreement (i.e., the scale means for system validity and fairness) were rather moderate. Principal leadership was stronger in managerial, collegial, instructional, and moral terms. Faculty culture was stronger and morale was up. But School I, a similarly high growth school as School C, did not fit this pattern. In the areas of principal leadership and faculty culture, the school looked 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, 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 at 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 seemed 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 a threat and high pressure, probably because of the school's recent decline in API. Neither the 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 as high in educational quality and low in organizational effectiveness.

A comparison between School I and School A confounded our model the most. Both schools were very similar on almost all measures, yet one grew by 30 more points in the last 2 years. Both engaged with the accountability system constructively and perceived their principals as relatively weaker. Interview data revealed that the principal at School A (with 600 more students) was seen as less of a presence in instructional affairs compared with the higher growth School I, yet the faculty had a strong collective tradition preceding the current principal's tenure. Thus, uncertainty remained.

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 characteristics of organizational 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 were matched to a degree. Our fictitious traveling practitioners could learn some valuable lessons if they selected the right time frame. If they selected schools on the basis of absolute API score 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 the implementation of a structured language arts program, are more developed at schools that experienced recent growth on the state performance indicator. But they also would find schools that grew without this exceptional leadership and faculty culture and schools that have 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 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 linked systematically 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, even this outlier school, let alone the other high-API schools, cannot be clearly distinguished from a much lower performing school along educational quality criteria (besides API). A lack of systematic connection between absolute API performance with other educational quality measures seems to be grounds to question the practical relevance of the accountability system.

For growth, the picture looks better. Confounding individual cases notwithstanding, 2-year growth preceding data collection 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 presumably are 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 at 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 longer term growth.

A more positive attitude toward accountability could theoretically be a cause, coincidence, or result of higher growth. Performance success and positive attitudes toward the judging authority often go hand in hand and can mutually reinforce each other. A sense of pressure and threat seems to have developed at 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 at schools that had posted solid growth over 6 years, as could have been expected. Instead, it occurred at three schools at which principals had forged a consensus around accountability. Two of the three schools (C and I) had gone on the upswing very recently yet were still classified by the system as very low performing and on the verge of major corrective action. Teachers at these schools were still insecure about their future prospects, but nevertheless more positive about accountability.

Teachers at 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. School C, at which organizational culture and accountability variables are highly positive according to our models, exhibits this pattern most strongly. In other words, what teachers do to become better on the state tests may not 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 at the successful schools, teachers do accomplish to improve student achievement as measured by standardized tests. How?

Our observations suggest that teachers at 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 the quality of students' writing.

A comparison between the top-API School B and the bottom-API School D illustrates what we mean. Both schools have been described in previous sections. Compared with 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 pattern as below average (for the nine case sample) *meaningfulness of accountability* and *instructional coherence* ratings. But teachers at this school did not teach any worse or provide 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 curricular 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 at this middle school and folded into the teaching of the literacy programs. The programs were implemented well, as reflected in an above-average proportion of engaging lessons. Doubts about the adequacy of science and social studies instruction were aired openly in this faculty, but for the time being, a collective commitment had been made to focus on remedial literacy.

Eight of the nine schools, the above-mentioned School D being the exception, followed in the footsteps of School B. But they did not implement their standards-aligned and structured programs with nearly as much enthusiasm nor to the exclusion of other subject matter, although at 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 through internalizing accountability. In others, the programs were in place but implemented with a more casual attitude. In neither case can the mere presence of a particular program 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 at 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.

Implications for System Design

Although we have never found a theoretical justification for a high-pressure approach to 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 fall-back solution" (Mintrop & Trujillo, 2005). On the other hand, we submit, as far as a study of this scale can, that it is the power of practically

meaningful aspects of accountability, combined with a supportive and open organizational climate and mild pressure, that drives schools to grow. This conclusion confirms findings from an earlier study (Mintrop, 2004). Even though successful schools in our case selection had a more positive attitude relative to less successful ones, educators at the nine schools overall had a fairly dim view of the system's meaningfulness for their work, as evidenced by the mild strength of agreement for scales such as validity, fairness, and raised expectations. This orientation 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. To foster the creation of better 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 and some by the school. This openness is impossible as long as main drivers of school improvement are school rankings based on presumably iron-clad performance indicators and the threat of sanctions.

Almost all nine schools in our selection are faced with sanctions, mostly as a result of having failed to meet the more stringent federal adequate yearly progress. Two of our three schools, identified as high growth in 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 (i.e., when they willingly ignore justified expectations). This is clearly not the case at these two high-growth schools or at others that were less successful with their strenuous efforts. Our case studies show that subtle patterns of takeoff, stalling, or coasting (Stoll & Fink, 1998), 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 the No Child Left Behind Act, 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.

Revisiting the Paradox

Practitioners assume a connection between high performance in the system and broader school quality, yet they claim to learn little from accountability success cases. Systems measuring students reward or punish the unmeasured actions of adults. How do we deal with these seeming contradictions in light of our findings? Our study shows that the nine schools, regardless of their success or failure in the accountability system, are surprisingly similar in many dimensions of quality and in the improvement strategies they use. So when our fictitious practitioners traveling across the schools say that "our kids" are behaving not that differently from "theirs" and "we" use the same programs that "they" do, they are probably right. Encompassing claims of overall quality differences are probably not warranted in light of our findings. Yet there are some differences that do apply, and practitioners can learn from them if they want to become more successful in the system, namely, the way higher growth schools engage with accountability. Successful schools seem to take the accountability system more seriously and seem more disposed to work on the intensive transmission of state-assessed knowledge, by no means a trivial pursuit, but one should recognize its limitation for what we consider school quality. Such dispositions and concomitant strategies help schools carefully align student and adult actions and hence to bridge the system's assessment and accountability functions. But an element of bafflement, even at the most successful schools in the system, holds sway: So many worthy adult initiatives, so much strenuous exertion to raise the quality of education for the schools' challenging clientele appears to go unrewarded. A principal

who is revered by the community for having brought peace, order, and civility to his middle school "counts for nothing" without his meeting API targets; an orchestra that was once the pride and glory of a struggling middle school slowly turns into a mere distraction; a school that shot up on one indicator (API), but misses the boat on yet another (adequate yearly progress), now fears moving up in program improvement status despite all-out improvement efforts. The disconnect between what adults in the schools consider worthy and therefore chance as a chip in the accountability race and what the system is able to measure and recognize remains indeed paradoxical.

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.

We need to remain aware of instabilities in the state's performance measures. Claims of connection between an accountability system and school quality depend on the quality of the state's accountability measures. The scholarly discussion about California's API is ongoing. Some have argued its statistical robustness, and others have found it to be flawed. To the degree that the API is a faulty measure, inconsistencies between API and school quality can be expected.

Our claims, restricted as they are, depend on an appropriate choice of measures of school 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 and that accountability systems are designed to measure not educational processes but student outcomes. 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; some lost funding or changed class size or turned over rapidly. These changes make claims of consistent relationships between indicated performance and realities on the ground rather heroic. Thus, we needed to avoid the "causality trap" yet search for possible connections that may or may not justify inferences that adults commonly make. 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 for the reader to cross-check our interpretations of the data. If we reinforced scholars' and practitioners' concern for the practical relevance of accountability systems with our nine cases in one state, we have reached our goal.

Appendix A
Teacher and Student Questionnaire Variables

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 students' needs
Test importance—personal	Students feel high state test scores are personally important
Test importance—whole school	Students feel high state test scores are 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
Faculty culture	
Collegiality	Cooperative effort and support among staff members
Pulling together	Cooperative effort and support among staff driven by accountability demands
Norms of performance	Teachers set and hold one another 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 because of 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

(continued)

Appendix A (continued)

Name	Definition
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 computers

Appendix B
Student Survey Scales

Item	Factor Loading
Student educational experience	
Academic engagement[a] (Cronbach's $\alpha = .69$)	
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
Academic press[a] (Cronbach's $\alpha = .77$)	
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
Teacher care[b] (Cronbach's $\alpha = .79$)	
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
Peer collaboration[b] (Cronbach's $\alpha = .74$)	
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
Safety[a] (Cronbach's $\alpha = .74$)	
How safe do you feel	
around the school?	.711
in the hallways and bathrooms of the school?	.678
in your classes?	.614

(continued)

Appendix B (continued)

Item	Factor Loading
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.	
Background	
Familial support[a] (Cronbach's $\alpha = .79$)	
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
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?	

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

Item	Factor Loading
Accountability	
Goal importance[a] (Cronbach's $\alpha = .76$)	
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.[b]	.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.[b]	.691
External validation (Cronbach's $\alpha = .81$)	
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
Authoritativeness (Cronbach's $\alpha = .72$)	
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
Threat (Cronbach's $\alpha = .89$)	
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
Focus[a] (Cronbach's $\alpha = .77$)	
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

(continued)

Appendix C (continued)

Item	Factor Loading
Diagnostics[c] (Cronbach's α = .88)	
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.[b]	.739
The state tests provide information that helps schools improve.	.875
State test results help identify students who need additional academic help.	.787
Validity[a] (Cronbach's α = .72)	
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
Fairness[a] (Cronbach's α = .75)	
For the most part, teachers are unfairly judged by the accountability system.[b]	.750
I resent being judged based on schoolwide test scores and the performance of other teachers.[b]	.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.[b]	.719
Our students are not behind because of the teachers they have, but because of the conditions in which they have to grow up.[b]	.760
Realism[a] (Cronbach's α = .79)	
The performance expectations of the state are for the most part unrealistic.[b]	.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.[b]	.713
The state assessments are unrealistic because too many tasks are too hard for our students.[b]	.688
Raised expectations (Cronbach's α = .88)	
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
Goal integrity[d]	
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	
Leadership	
Urgency (Cronbach's α = .73)	
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
Principal support[a] (Cronbach's α = .90)	
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

(continued)

Appendix C (continued)

Item	Factor Loading
Principal control[a] (Cronbach's α = .64)	
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
School management (Cronbach's α = .93)	
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.[b]	.832
Open communication (Cronbach's α = .86)	
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
Autonomy (Cronbach's α = .81)	
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
Moral leadership (*r* = .75)	
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.	
Instructional leadership[e] (Cronbach's α = .91)	
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
Faculty culture	
Collegiality[f] (Cronbach's α = .81)	
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
Pulling together (Cronbach's α = .80)	
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
Norms of performance (Cronbach's α = .90)	
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

(continued)

Appendix C (continued)

Item	Factor Loading
Learning orientation[f] (Cronbach's α = .76)	
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.[b]	.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
Satisfaction (r = .52)	
How often do you feel satisfied	
with your work as a teacher?	
with your school overall?	
Efficacy and qualifications	
Instructional efficacy (Cronbach's α = .75)	
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.[b]	.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.[b]	.609
For the most part, my students are engaged in my lessons.	.730
Test-related efficacy (r = .52)	
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.	
Colleagues' skills[a] (Cronbach's α = .75)	
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.[b]	.778
Change strategies	
Program coherence[e] (Cronbach's α = .81)	
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.[b]	.777
Many special programs come and go at this school.[b]	.831
You can see real continuity from one program to another at this school.	.810
Strategic orientation (r = .61)	
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.	
Data usage (Cronbach's α = .87)	
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

(continued)

Appendix C (continued)

Item	Factor Loading
District operational system[g] (Cronbach's α = .87)	
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
District instructional system (Cronbach's α = .75)	
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
Background	
Parental support[e] (Cronbach's α = .83)	
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

Note. API = Academic Performance Index; AYP = adequate yearly progress.
a. Adapted from Mintrop (2004).
b. Values are reversed.
c. Adapted from Bomotti, Ginsberg, and Cobb (2002).
d. Scores calculated on the basis of differences between like items.
e. Adapted from Consortium on Chicago School Research (2003b).
f. Adapted from McLaughlin and Talbert (1993).
g. Adapted from SRI International, Policy Studies Associates, and Consortium for Policy Research in Education (2003).

Appendix D
Classroom Observation Measures

Measure	Definition
	Percentage of snapshots in which
Noninstructional time	Classroom activity was not related to student learning
Time on task	At least three quarters 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 used 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)

Appendix E
Demographic Characteristics of the Nine Selected Cases Over Time, 1999–2005

School		1999	2000	2001	2002	2003	2004	2005
					English learners			
High	B	32	32	29	28	29	26	28
	E	25	23	22	23	21	23	18
	A	36	33	36	35	40	35.8	43
	G	25	9	35	43	38	38	31
	H	45	49	44	46	49	49	44
Low	C	35	38	35	35	32	28	26
	I	40	40	40	51	51	42	39
	D	24	31	31	21	20	22	22
	F	36	40	35	37	39	24	29
					Free or reduced-price lunch			
High	B	83	81	78	79	78	76	78
	E	65	54	60	60	68	67	69
	A	79	78	77	79	79	79.9	83
	G	77	76	77	81	63	76	85
	H	80	81	87	70	85	88	77
Low	C	87	88	86	85	89	100	100
	I	87	85	83	89	90	100	100
	D	52	51	49	51	59	59	59
	F	92	95	98	97	96	96	97
					Parent education[a]			
High	B	2.09	2.21	2.23	2.03	2.19	2.04	2.03
	E	2.27	2.21	2.24	2.13	2.14	2.14	2.18
	A	2.37	2.39	2.42	2.32	2.21	2.11	2.09
	G	2.11	2.56	1.86	1.7	1.96	1.88	2.02
	H	1.82	1.74	1.7	1.71	1.71	1.73	1.81
Low	C	2.02	1.99	2.28	2.39	2.38	2.26	2.25
	I	2.07	2.26	2.29	2.11	1.97	1.96	2.09
	D	2.29	2.36	2.38	1.57	2.21	2.15	2.13
	F	1.86	n/a	2.06	1.89	1.93	1.84	1.81

Source. California Department of Education (2006).

a. 1 = *not a high school graduate*, 5 = *graduate school*. Parent education is subject to the inaccuracies of self-reported data.

Notes

[1]The ground for this work was prepared during an earlier study directed by one of the authors (Mintrop, 2003, 2004).

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

[3]Core variables for this study can be found in a Center for Research on Evaluation, Standards and Student Testing (CRESST) technical report (Mintrop & Trujillo, 2007) and in the online appendix for this article.

[4]Here we briefly describe the properties of our instruments and the ways they were administered. For a more in-depth discussion, including the statistical properties of all survey scales, interrater agreement calculations, descriptive and inferential statistics, and detailed data collection and analysis procedures, refer to the CRESST technical report.

The student questionnaire consisted of 50 items capturing students' perceptions of quality as well as family background, awareness of accountability, and test-taking attitudes. It was designed using items from previously conducted student surveys (Consortium on Chicago School Research, 2003a; International Association for the Evaluation of Educational Achievement, 2003; Organisation for Economic Co-operation and Development, 2000) and newly designed, pilot-tested items. 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 over- or undersampling with weights. The overall response rate was 96% (between 94% and 99% across the nine schools).

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. Almost all lessons were observed by two observers who were trained extensively at pilot schools. 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 classes in each curricular track. Throughout each lesson, we rated three 5-minute snapshots spaced evenly throughout the observation. The classroom observations were followed by a postobservation interview in which we tried to ascertain how teachers had approached planning and whether the observed lessons were tied to possible strategies of instructional improvement. Finally, we wrote a descriptive summary of each lesson according to a specified observation guide.

Student writing samples were collected from English language arts classes at each school. Like the student questionnaire, we sampled the 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-, medium-, and low-quality exemplar. We collected 390 pieces of writing from 130 classes. As with the student questionnaire data, we adjusted for slight over- 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, interrater agreement of 90% on the 20% of the sample overlapping between the two raters was achieved.

The teacher questionnaire consisted of over 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 came from a variety of sources (Author, 2004; Consortium on Chicago School Research, 2003b; McLaughlin & Talbert, 1993; 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. This questionnaire was administered to all teachers in the nine schools. The 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. One hundred fifty-one 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 article. In addition, we collected data on the schools' demographic backgrounds, school conditions, and inputs.

[5]We used data from the California Basic Educational Data System (California Department of Education, 2005), which contains annual Academic Performance Index (API) score data as well as information on the 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 or junior high schools were selected that had complete records of 4 years of API scores and demographic information. We predicted API scores on the basis of the School Characteristics Index, which is a composite index of the demographic characteristics (i.e., percentage of pupils with free or reduced-price lunch (FRPL) participation, percentage of English language learners, ethnic background, and student mobility) and a proxy for school capacity (i.e., percentage of teachers with full credentials). The School Characteristics Index is a variable contained in the state database. Because student populations and the basis for the API changed over the years (e.g., shifting from norm-referenced to criterion-referenced 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. 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. The API scores on which we based our case selection were as follows:

	Low				High				
	F	D	I	C	H	G	A	E	B
1999 API score	478	503	478	481	442	521	489	523	445
2003 API score	537	577	533	548	595	657	616	634	605
Score difference	59	74	55	67	153	136	127	111	160

[6]Classroom observation measures were defined as the percentage of snapshots in which each of the following was observed. *Noninstructional time:* classroom activity was not related to student learning; *time on task:* at least three quarters 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 used active instructional techniques (e.g., modeling, coaching, recitation, discussion, assessment); and *cognitive complexity:* students engaged in cognitively demanding activities (e.g., demonstrate/explain, analyze/investigate, evaluate, generate/create).

[7]Our original sampling and classification were based on 4 years' growth on the API between 1999 and 2003. On the basis of those years, schools in our two performance groups differed nicely both on absolute and relative measures. The two top- and bottom-growth schools differed by 100 API points (168 points by 2005), whereas 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. This difference had 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 ranks (i.e., API performance decile), whereas all of the schools in our high group had grown at least by one rank.

[8]Reducing the comparison from nine to five schools (three top-API-growth schools vs. two bottom-API-growth schools over 6 years), thus increasing the marginal difference between the two groups, made three accountability-related measures (focus, validity, and fairness) significant at $p < .05$, but this is still far from the kind of more encompassing patterns that could establish practical relevance.

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*(3), 421–455.

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

Baker, E., & Linn, R. (2004). Validity issues for accountability systems. In S. Furhman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47–72). New York: Teachers College Press.

Baker, E., Goldschmidt, P., Martinez, F., & Swigert, S. (2002). *In search of school quality and accountability: Moving beyond the California Performance Index (API)*. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

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

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)*. Available at http://www.cde.ca.gov/ds/sd/cb/

California Department of Education. (2006). *API description: Overview of the Academic Performance Index (API)*. Available at http://www.cde.ca.gov/ta/ac/ap/apidescription.asp

Carlson, D. (2006). *Focusing state educational accountability systems: Four methods of judging school quality and progress*. Dover, NH: Center for Assessment.

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. Department 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.

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

Goe, L. (2004). *An evaluation of California's immediate intervention/underperforming schools program*. Unpublished doctoral dissertation, University of California, Berkeley.

Goe, L. (2006, April). *An evaluation of the Immediate Intervention/Underperforming Schools (II/USP): 1999–2004*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

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: Routledge Falmer.

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.

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

Hanushek, E. (1994). *Making schools work: Improving performance and controlling costs*. Washington, DC: Brookings Institution.

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.

Hill, R. (2001). *The reliability of California's API*. Dover, NH: The Center for Assessment.

Ingersoll, R. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499–534.

International Association for the Evaluation of Educational Achievement. (2003). *TIMSS 2003 main survey, student questionnaire, Grade 4*. Amsterdam, the Netherlands: Author.

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

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

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*(3), 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*(2), 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, the Netherlands: Swets & Zeitlinger.

Malen, B., & Muncey, D. (2000). Creating "a new set of givens"? The impact of state activism on school autonomy. In N. D. Theobald & B. Malen (Eds.), *Balancing local control and state responsibility for K–12 education* (pp.199–244). Larchmont, NY: Eye on Education.

McBeath, J., & Mortimore, P. (Eds.). (2001). *Improving school effectiveness*. Buckingham, UK: 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.

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

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 November 2, 2005, 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*(4), 275–300.

Mintrop, H., & Trujillo, T. (2005). Corrective action in low-performing schools: Lessons for NCLB implementation from state and district strategies in first-generation accountability systems. *Educational Policy Analysis Archives*, *13*(48). Available at http://epaa.asu.edu/epaa/v13n48.html

Mintrop, H., & Trujillo, T. (2007). *School improvement under test-driven accountability: A comparison of high and low performing middle schools in California* (CSE Tech. Rep. No. 717). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

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: 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*(4), 297–321.

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

O'Day, J. (2002). Complexity, accountability, and school improvement. *Harvard Education Review*, *72*(3), 293–329.

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

Organisation for Economic Co-operation and Development. (2000). *PISA 2000 technical report*. Paris: Author.

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

Rogosa, D. (2003). *Confusions and consistency in improvement*. Retrieved March 13, 2007, from http://www-stat.stanford.edu/%7Erag/api/consist.pdf

Rogosa, D., & Haertel, D. (2003*). Deceived and confused: An attempt to reconcile the numbers in the public forum on school accountability report, "A better student data system for California."* Available at http://www-stat.stanford.edu/~rag/api/Deceive.pdf

Rowan, B., Chiang, F., & Miller, R. (1997). Using research on employees' performance to study the effects of teachers on student's achievement. *Sociology of Education*, *70*(4), 256–284.

Russell, M. (2002). *California's accountability system and the API, expert witness report for Eliezer Williams et al. v. State of California*. Available at http://www.decentschools.org/expert_reports/russell_report.pdf

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

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

Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford, UK: Pergamon.

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

Springboard Schools. (2005). *Challenged schools, remarkable results: Three lessons from California's highest achieving high schools*. San Francisco: 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: U.S. Department of Education, Planning and Evaluation Service.

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* (pp. 189–206). London: Falmer.

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.

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: Author.

Williams, T., Kirst, M., Haertel, E., et al. (2005). *Similar students, different results: Why do some schools do better? A large scale survey of California elementary schools serving low-income students.* Mountain View, CA: EdSource.

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

Authors

HEINRICH MINTROP is an associate professor at the Graduate School of Education and director of the Doctoral Program in Educational Leadership at the University of California, Berkeley, 3647 Tolman Hall, Berkeley, CA 94720; mintrop@berkeley.edu. His areas of specialization are school improvement, test-driven accountability systems, and democratization in education. He holds an MA in political science and German studies from Free University, Berlin, and a PhD in education from Stanford University.

TINA TRUJILLO is a doctoral candidate in urban schooling at the Graduate School of Education and Information Studies, University of California, Los Angeles, Moore Hall, Box 951521, 405 Hilgard Avenue, Los Angeles, CA 90095-1521; tina.trujillo@ucla.edu. Her current research interests include urban district reform, high-stakes accountability policies, and school improvement. She holds an MA in educational foundations, policy, and practice and a BA in political science, both from the University of Colorado at Boulder.

Manuscript received July 28, 2006
Final revision received August 21, 2007
Accepted August 31, 2007