Chapter 2
School Leadership, Evidence-Based Decision Making, and Large-Scale Student Assessment

Kenneth Leithwood

Current preoccupations with student assessment data have their roots in the demanding external accountability policies imposed on schools beginning in the early to mid 1990s. These policies have dramatically reshaped the responsibilities of school leaders. In almost all contemporary school systems those in formal leadership roles, particularly principals, are now held directly accountable for improvements in the performance of their students. The consequences of failure for principals range from mildly negative (pressure from district administrators, dissatisfaction from some parents) to career threatening (removal from the position), depending on the specific policy and district context in which principals find themselves. Large-scale assessment results are the primary, often the only, instruments used for such accountability purposes. It should not be surprising, then, to find school leaders looking to the results of such tests for clues to assist them in their school improvement task. Just how likely is it, however, that large-scale assessment data will provide such clues?

Typically the brainchild of policy makers, large-scale assessments, according to McDonnell's (2005) review, are expected to accomplish a wide range of purposes. In addition to the accountability they impose on principals and other educators, policy makers expect large-scale assessments to:

- Provide current status information about the system.
- Help with instructional decisions about individual students.
- Bring greater curricular coherence to the system.
- Motivate students to perform better and parents to demand higher performance.
- Act as a lever to change instructional content and strategies.
- Certify the achievement or mastery of individual students.

K. Leithwood (✉)
Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada
e-mail: kenneth.leithwood@utoronto.ca

C.F. Webber and J.L. Lupart (eds.), *Leading Student Assessment*,
Studies in Educational Leadership 15, DOI 10.1007/978-94-007-1727-5_2,

This is an astonishingly broad range of purposes, several of which are near and dear to the hearts of improvement-minded school leaders (e.g., help with instructional decisions about individual students). But as McDonnell (2005) pointed out, efforts to use the same test for multiple purposes will often run afoul of standards considered vital by the measurement community. Invoking such standards established by the American Educational Research Association (2000), McDonnell described,

> the need to base high stakes decisions on more than a single test; validated tests for each separate use; and [provision of] adequate resources for students to learn the content being tested (p. 46).

Very few large-scale assessments come close to meeting these standards, an inauspicious point of departure for considering their use by school leaders.

While current expectations of school leaders roughly parallel the expectations of leaders in most organizations—"improve the bottom line"—we need to acknowledge the uniquely demanding nature of this mission for school leaders. Student and family background variables are widely believed to account for more than half of the variation across schools in student achievement. Indeed, the best evidence currently available (e.g., Creemers and Reetzig 1996) suggests that of all variables within the school, those ostensibly under the direct control of principals collectively explain 12–25% of the variation in student achievement.

The implication of this evidence about the proportion of student achievement, explained by what schools do directly, indicates that successfully improving student performance depends on school leaders exercising enormous leverage over the variables which they are able to influence, leverage likely to depend on exceptionally sensitive information about students' learning and how it might be improved. What do school leaders need to know to be successful in improving student performance in their schools? Is large-scale assessment information about the performance of their students sufficient? If not, what else would be helpful? Just how large a proportion of what leaders need to know is captured by large-scale assessment information about their students' performance?

This chapter grapples with these questions as a means of widening the conversations, now underway in many educational jurisdictions, about evidence-based decision making (Earl 2001), what this means for principals and for their role in advocating for evidence-based decisions on the part of their school colleagues. The argument I develop in the chapter is as follows:

> Evidence about student performance is clearly essential for school leaders to successfully carry out their school improvement task. But such information provided by large-scale assessment is often fraught with limitations and is always woefully insufficient. Actually improving student performance also requires information of a very different order, and the absence of this information in most schools greatly diminishes school leaders' chances of success.

This argument is developed through the examination of five challenges facing school leaders in their efforts to improve the performance of their students:

- Compensating for the critical limitations of large-scale assessment data in determining the current status of student learning.
- Estimating progress and sustaining continuous improvement.

- Responding to the absence of robust evidence about the causes of students' current performance.
- Improving the organizational conditions that support productive school improvement efforts.
- Overcoming common errors in human judgment.

Challenge One: Compensating for the Critical Limitations of Large-Scale Assessment Data in Determining the Current Status of Student Learning

What are the challenges for leaders attempting to use such measures? Typically part of national, provincial, or district student testing programs, these measures have three well-known limitations for leaders: A narrow focus, unknown local reliability, and significant delays in the reporting of results.

Narrow Focus

Most large-scale testing programs confine their focus to math and language achievement with occasional forays into science. Only in relatively rare cases have efforts been made to test pupils in most areas of the curriculum, not to mention cross-curricular areas such as problem solving (see Saskatchewan for an exception on this) or teamwork. Technical measurement challenges, lack of resources, and concerns about the amount of time for testing explain this typically narrow focus of large-scale testing programs.

This means, however, that evidence of a school's impact on student achievement using these sources is evidence of effects on pupils' literacy and numeracy. Even when testing programs are expanded, as in the case of Alberta's addition of science and social studies in Grade 9, they almost never come close to reflecting the full range of important cognitive, social, and affective goals reflected in the school system's curriculum policies. So the contribution of schools to individual students and to society, more generally, is always underestimated by these large-scale testing programs, typically by a huge margin. The consequences of such underestimation are much more than technical. Such consequences include the potentially serious erosion of parent and community support for public schooling, an erosion felt first at the local school level—what gets measured gets valued.

The implication of this challenge for school leaders is twofold. First, school leaders will need to adopt additional assessments in their schools, assessments designed to measure a larger proportion of the goals their schools aim to develop among their students. Leaders moving in this direction, second, will need to assume an educative role with parents and community members, helping them to appreciate the value of this larger range of goals for their children.

Lack of Reliability at the Local Level

Lack of reliability at the school and classroom level is a second limitation of many large-scale testing programs. Most such programs are designed to provide reliable results only for quite large groups of pupils; results aggregated to national, state, or district levels are likely to be reliable (Heubert and Hauser 1999). This also may be the case for the aggregated results of student performance in relatively large schools, but not for performance results disaggregated by student groups or categories as required in the US *No Child Left Behind* legislation. As data are disaggregated, or the number of pupils diminishes, as in the case of a classroom, single school, or even a small district or region, few testing systems claim to even know how reliable are their results (e.g., Wolfe et al. 2004).

The likelihood, however, is that they are almost certainly not very reliable, thereby challenging their diagnostic value to school leaders and their staffs. Indeed Wilson (2004) has argued that the testing community has paid insufficient attention to the central place of classrooms in assessing student performance. This lack of reliability warrants restricting the analysis of large-scale assessment results to data aggregated above the level of the individual school or leader—in direct opposition to virtually all the systems of school reporting in most British, Canadian, and American contexts.

Delays in Reporting Results

To the three technical limitations of large-scale testing programs discussed above, most school leaders would quickly add "lack of timely reporting of results." It is by no means unreasonable for teachers and principals to complain that the test performance of students, collected in the spring of the year but not made available to them until the fall, lack diagnostic currency for their instructional and school improvement purposes. While many jurisdictions now aim to reduce this reporting lag, it remains a common problem.

For instructional and school improvement purposes, testing ideally would occur in the very early fall with results available to schools by mid-fall, at the latest. The fact that this is unlikely to happen anytime soon simply reinforces the value of schools adopting their own measures of student achievement, in addition to participating in large-scale assessments.

School leaders wanting reliable evidence about student performance in their schools and classrooms will usually have to supplement large-scale test results by administering additional (valid) assessments known to provide reliable estimates of individual student performance. They would also be advised to educate their staffs and communities about the reasons for selecting measures of achievement in addition to those provided by large-scale assessments.

Challenge Two: Estimating Progress and Sustaining Continuous Improvement

Monitoring the extent to which a school improves the achievement of its pupils over time is a much better reflection of a school's and a leader's effectiveness than is the school's annual mean achievement scores. Many educational systems now acknowledge this to be the case, even though year-over-year comparisons have been the norm until recently, as illustrated by the "adequate yearly progress" targets established for most principals working in compliance with the US *No Child Left Behind* legislation. Technically speaking, however, arriving at a defensible estimate of such change is difficult. Simply attributing the difference between the mean achievement scores of this year's and last year's pupils on the province's literacy test to changes in a school's effectiveness overlooks a host of other possible explanations:

- Cohort differences: This year's pupils may be significantly more or less advanced in their literacy capacities when they entered the cohort. Such cohort differences are quite common.
- Test differences: While most large-scale assessment programs take pains to ensure equivalency of test difficulty from year to year, this is an imperfect process and there are often subtle and not-so subtle adjustments in the tests that can amount to unanticipated but significant differences in scores.
- Differences in test conditions: Teachers are almost always in charge of administering the tests and their class's results on last year's tests may well influence the nature of how they administer this year's test (more or less leniently) even within the guidelines offered by the testing agency.
- External environment differences: Perhaps the weather this winter was more severe than last winter and pupils ended up with six more snow days—six fewer days of instruction—or a teacher left half way through the year, or was sick for a significant time.
- Regression to the mean: This term is used by statisticians to capture the highly predictable tendency for extreme scores on one test administration to change in the direction of the mean performance on a second administration. So schools scoring either very low or very high in a year can be expected to score extremely less during the next, quite aside from anything else that might be different.

To demonstrate the powerful effects that these and related factors have on a school's performance over time, my colleagues and I have recently examined the achievement trajectories of all elementary schools in Ontario for which Educational Quality and Accountability Office (EQAO) Grade 3 reading scores were available (we will be undertaking the same examination of scores in other content areas, as well). We examined the achievement trajectories of these schools over three annual testing cycles (2004–2005, 2005–2006, and 2006–2007). For each of these years, a school's performance could stay unchanged or stable (S), decline (D), or increase (I). A "continuous improvement" trajectory would, of course, consist of increased achievement over each year for 3 years (III). Results are summarized in Table 2.1.

Table 2.1 Patterns of school performance in Ontario schools (2004–2007)

Pattern of change (3 years)[a]	Number of schools	Percent of schools
DSS	1	.0
DSD	7	.2
DSI	29	.8
DDS	5	.1
DDD	23	.6
DDI	175	4.6
DIS	33	.9
DID	341	8.9
DII	353	9.2
ISS	3	.1
ISD	14	.4
ISI	27	.7
IDS	17	.4
IDD	143	3.7
IDI	536	14.0
IIS	28	.7
IID	407	10.7
III	284	7.4
Total	2,494	65.3

[a] *S* stable/unchanged; *D* decreased; *I* increased

The left column of the table indicates that 18 possible achievement patterns were actually found among the schools in the province. The next column of Table 2.1 indicates that no single pattern of the 18 possibilities represented more than 14% of the schools (IDI) and only 7.4% of schools in the province demonstrated a continuous improvement pattern (III). The other most common patterns were IID (10.7%), DII (9.2%), and DID (8.9%). Combining the results for the two most positive patterns (DII and III) captures about 16% of the province's schools.

To add further meaning to these results, individual patterns of achievement have been clustered into six broader trajectories:

- *Temporary success*: This broad pattern consists of an increase in student performance the first year (2004–2005) followed by 2 years of either no change (S) or decreased performance (ISS, IDD, IDS, and ISD). This pattern encompassed 4.6% of the province's schools.
- *Temporary failure*: This pattern consists of a decrease in performance in the first year followed by stable or increased performance over the next 2 years (DIS, DSI, DII). About 11% of the schools fell into this pattern.
- *Longer term success*: Keeping in mind that we are working with only 4 years of data (three annual cycles of change), this pattern consisted of increased performance the first year followed by steady or increased performance in the subsequent 2 years (III, ISI, IIS). Approximately 9% of schools reflected this pattern.
- *Longer-term failure*: Schools demonstrating this pattern had decreased performance the first year followed by 2 years of either stable or decreased performance. Fewer than 1% of the province's schools fit this pattern.

- *No predictable direction*: The performance of almost a quarter (23%) of schools in the province demonstrated this pattern (DID and IDI).
- *Running out of steam*: More than 10% of the province's schools fit this pattern, one that entailed 2 years of increased performance followed by a third year of decline (IID).
- *Finally catching a break*: A pattern consisting of 2 years of decline followed by a year of increased performance described the achievement trajectory in 4.6% of the schools in the province (SSI, DSI, DDI, and SDI).

These results indicate quite clearly that when large-scale assessment results are the criterion by which school improvement is measured, such improvement is a very bumpy road for most schools. *No predictable direction* is the pattern of change evident for by far the largest group of schools. *Long-term success* is a pattern evident among only 9% of schools, with the most desirable subpattern, "continuous improvement" (III), reflected in just 7.4% of the cases.

Linn (2003) has demonstrated with data from other educational jurisdictions that the challenges in estimating change from cross-sectional data, so clearly illustrated with our Ontario data, become less severe as change is traced over 3 or 4 years. It is the conclusions drawn from simply comparing this year's and last year's scores that are especially open to misinterpretation. While our Ontario results support (weakly) Linn's advice on this matter, they also suggest that even a longer time-frame horizon may provide conflicting inferences about the direction of student performance in a school.

The lesson for school leaders here is at least to focus on long-term trends (3, 4, or more years) in their schools' performance and not to be especially impressed or alarmed with changes in annual performance. Current efforts to develop systems for tracking the progress of individual students throughout their school careers could go a long way toward assisting school leaders in estimating the effects of their efforts to improve student achievement and compensating for the erratic long-term trends reflected in the Ontario data. School leaders should also be aware that, rhetoric aside, the concept of "continuous improvement," measured through the use of large-scale assessment results, is a rarely reached goal, and when it does appear it may not be of their own making, anyway.

Challenge Three: Responding to the Absence of Robust Information About the Causes of Students' Current Performances

Let's temporarily assume that the challenges described above have been addressed in some fashion. At minimum, for example, the report of large-scale assessment results provided by the province to the school is reliable and sufficiently broad to reflect the school's priorities for teaching and learning. Now the school has a reasonable estimate of the status of student learning in key areas of the curriculum. In some of these areas, students seem to be doing very well indeed, but there is clearly room for improvement in others. Perhaps, for example, the reading comprehension scores of

Grade 6 students are significantly below both the district mean as well as the level achieved by students in other schools in the district serving similar populations.

What now? To explain this third challenge for principals and teachers, I adopt a view of school improvement as a "problem" defined along the lines suggested by cognitive psychologists (e.g., Fredericksen 1984; Gagné 1985), as:

- A current state: For the school staff, this state is at least partly addressed by student assessment data.
- A goal state: This state is often clarified through some school improvement planning process in which the aspirations for students are clarified and some consensus among stakeholders on their importance is reached.
- Operators or solution paths: Strategies for transforming the current state into the goal state, likely to be ambiguous in the case of most school improvement problems.

The value or impact of school improvement solution paths selected by school staffs will typically depend on how well they account for the causes of their students' current achievement status. A recent case study by Timperley (2005) nicely illustrates how this element of problem solving was addressed in one elementary school. This case likely captures a form of "best practice" in comparison with other current approaches to this element of school improvement problem solving. Over several years, those leading the literacy initiatives in this school moved from engaging groups of teachers in discussions of assessment results, aggregated at the school and classroom levels, to discussions of such results disaggregated to the individual student level.

Timperley (2005) found dramatically different causal explanations for inadequate results invoked by teachers under each of these two conditions. High levels of data aggregation were associated with external-to-the-school teacher explanations (e.g., the children are from poor families and receive little support for literacy development in the home). Reports of individual student results were associated with much more reflection by teachers about their own instructional practices and how those practices should be altered.

Most of us will agree that such a shift in teachers' causal musings is a good thing. At least these teachers began to consider what they might do to improve their students' literacy skills rather than viewing the development of such skills as beyond their control. But the value of these teachers' reflections on their own practices depended entirely on their own sense-making capacities, the accuracy of their knowledge about their students, the sophistication of their own knowledge about how literacy develops, and the nature of effective literacy instruction. We might reasonably assume, under these circumstances, that the outcome would be highly variable across any group of teachers, a variability that might be reduced in the context of a collaborative school culture. But we should also expect a high degree of variability across schools with collaborative cultures because of significant differences in the collective instructional expertise of staffs.

The central point of this discussion is that, whether working in schools with isolated or collaborative cultures, teachers and school leaders almost always have to rely on their own often rich, but highly personal, situation-bound, and unsystematic evidence to explain the causes of the student achievement data with which they are confronted. And as the evidence synthesized by Nisbett and Ross (1980) many years ago indicates, peoples' causal explanations cannot extend beyond what people already know.

Responding usefully to this challenge entails working with staffs in a much more deliberate manner to surface the underlying social, affective, and cognitive causes of individual student achievement. The results of such activity will be a different order of evidence than the formative evidence suggested by Black and Wiliam (2004), for example. While school leaders often invite consultation around their school-improvement processes, such consultation rarely consists of help in diagnosing the challenges individual students are facing in improving their own learning. Such information, however, could serve a pivotal role in determining much of what goes into a school improvement plan.

Challenge Four: Improving the Organizational Conditions That Support Productive School Improvement

Direct and Indirect Approaches to Improvement

School leaders and their colleagues work at increasing the performance of their students in two distinct ways. These two routes to improvement are reflected reasonably well in the literatures now associated with *school improvement*, on the one hand, and *school effectiveness*, on the other (see Creemers 2007 for a recent explanation of this distinction). The school improvement literature concerns itself with often carefully planned processes intentionally aimed at accomplishing more or less specific outcomes. In this literature, school leaders typically occupy the foreground of the action and are portrayed as responsible for ensuring the development and implementation of school improvement processes (Silins and Mulford 2007).

The effective schools literature describes features of the school organization associated with greater than average impacts on student achievement. In this literature, "strong" school leadership is one of from 5 to 12 "correlates" of schools whose students perform beyond expectation (Teddlie and Stringfield 2007). While school leaders are not relegated to the background in this literature, their importance is balanced with the influence of at least a handful of other organizational features or conditions such as "safe and orderly culture" and "high expectations for student achievement" (e.g., Sackney 2007).

The second challenge described above reflects the planned and goal-driven nature of efforts which are the focus of the school improvement literature. The challenge taken up in this section is more reflective of the effective schools literature. Improving the organizational conditions that support school improvement acknowledges the often indirect nature of leadership effects and aims to build an organization in which powerful practices are nurtured in both explicit and quite subtle ways. This means creating conditions in the school which increase the likelihood that staffs will have both the will and skill to engage in effective practice, irrespective of intentional direction and action on the part of school leaders.

A large proportion of leadership effects research is conducted from this perspective, with promising conditions for improving student learning, assuming the role of "mediators"; these are conditions over which leaders have some direct influence and which, in turn, have a direct and significant effect on what and how well students learn. Such research assumes that "leadership" entails the exercise of influence, as reflected, for example, in a widely accepted definition of leadership:

> the ability of an individual to influence, motivate, and enable others to contribute toward the effectiveness and success of the organizations of which they are members (House et al. 2004, p. 15).

The Indirect Approach Illustrated

The challenge facing school leaders when they are working on the improvement of their schools in this indirect way is to identify the most promising "links in the chain" connecting what they do (their influence) to the performance of students. Figure 2.1 illustrates the way in which leadership effects on student learning have often been explored over the past two decades (e.g., Leithwood et al. 2006, 2004). This figure acknowledges the assumption, alluded to above, that leadership effects on students are largely, though not exclusively, indirect (Hallinger and Heck 1996; Pitner 1988). Leaders' indirect effects depend on the extent and nature of their influence on key variables or conditions (the mediators in Fig. 2.1) that are alterable through the direct intervention of leaders and which themselves have the power to directly influence the learning experiences of students.

Leaders' effects not only depend on the influence they exercise on these mediating variables, however. Such effects are either dampened or enhanced by what are often considered to be context variables (the moderators in Fig. 2.1). Student socioeconomic status (SES) is often used as a proxy for these variables. But specific features of students' family cultures, such as parental expectations for student success, respect for teachers, provision of a supportive environment in the home for school work, adequate nutrition, and the like, are the conditions that shape the social and intellectual "capital" which students bring to school and determine so much of what they learn (Walberg 1984).

School leader practices have been described in many ways. By way of illustration, however, there is considerable evidence at this point suggesting that the repertoire of almost all successful school leaders includes sets of specific practices aimed at establishing directions for their organizations, helping their colleagues build the capacities they will need to accomplish those directions, continually refining the design of the organization to facilitate teachers' work, and managing the instructional program in their schools (Leithwood and Riehl 2005; Leithwood et al. 2006).

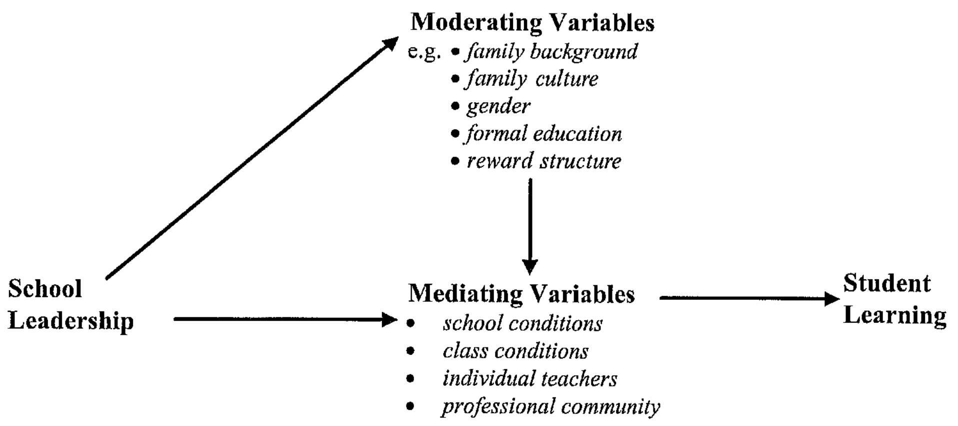

Fig. 2.1 How school leadership influences student learning

Figure 2.1 points in a general way at some of the most critical knowledge school leaders need, beyond knowledge of the status of their students' learning, if they are to be successful in improving such learning. This is knowledge about both mediators and moderators of their influence. Leithwood and Levin's (2005) recent review of leadership research identified some 11 school-level mediators, 6 classroom-level mediators, and 11 mediators concerned with teachers that have been included in recent leadership effects research. For illustrative purposes, the remainder of this section discusses three of these mediators of leadership effects on student learning—academic press, trust, and collective teacher efficacy. For each mediator, the evidence which would justify leaders' attention to it is summarized, as well as the advice to leaders provided by such evidence about how to increase the impact of that mediator on student learning.

Academic Press

This is a school-level mediator. In schools with significant degrees of academic press, administrators and teachers set high but achievable school goals and classroom academic standards. They believe in the capacity of their students to achieve and encourage their students to respect and pursue academic success. School administrators supply resources, provide structures, and exert leadership influence. Teachers make appropriately challenging academic demands and provide quality instruction to attain these goals. Students value these goals, respond positively, and work hard to meet the challenge.

Research on effective schools identified academic press as one important correlate of effective school climate and linked it with student achievement as early as the late 1970s and early 1980s. Of the more than 20 empirical studies of academic press which have been published since about 1989, by far the majority have reported significant, positive, though moderate relationships between academic press and student achievement, most often in the area of math, but extending to other subjects such as writing, science, reading, and language, as well. academic press is described as explaining almost 50% of the between-school variability in mathematics and reading in Goddard, Sweetland, and Hoy's (2000) study, for example, after controlling for the effects of students' family backgrounds. Most of the evidence suggests that a school's academic press makes an especially valuable contribution to the achievement of disadvantaged children.

Academic press is one of the more malleable conditions for leadership intervention and a small number of studies have provided some guidance on the practices likely to increase a school's academic press (e.g., Alig-Mielcarek 2003; Jacob 2004; Jurewicz 2004). Included among those practices are the following:

- Promoting school-wide professional development.
- Monitoring and providing feedback on the teaching and learning processes.
- Developing and communicating shared goals.
- Being open, supportive, and friendly.
- Establishing high expectations.
- Not burdening teachers with bureaucratic tasks and busy work.
- Helping to clarify shared goals about academic achievement.
- Grouping students using methods that convey academic expectations.
- Protecting instructional time.
- Providing an orderly environment.
- Establishing clear homework policies.
- Monitoring student performance in relation to instructional objectives.
- Base remediation efforts on the common instructional framework.
- Requiring student progress reports to be sent to the parents.
- Making promotion dependent on student mastery of basic grade level skills.

Teacher Trust in Colleagues, Parents, and Students

An individual-level teacher mediator, trust is conceptualized in many different specific ways (e.g., Dirks and Ferrin 2002). But almost all efforts to clarify the nature of trust include a belief or expectation, in this case on the part of most teachers, that their colleagues, students, and parents support the school's goals for student learning and will reliably work toward achieving those goals. Transparency, competence, benevolence, and reliability are among the qualities persuading others that a person is trustworthy. Teacher trust is critical to the success of school improvement initiatives and nurturing trusting relationships with students and parents is a key element in improving student learning. (e.g., Lee and Croninger 1994).

Trust remains a strong predictor of student achievement even after the effects of student background, prior achievement, race, and gender have been taken into account in some recent studies of trust in schools. Goddard (2003) argued that when teacher–parent and teacher–student relationships are characterized by trust, academically supportive norms and social relations have the potential to move students toward academic success. Results of a second study by Goddard and his colleagues (2001) provide one of the largest estimates of trust effects on student learning. In this study, trust explained 81% of the variation between schools in students' math and reading achievement.

Principal leadership has been highlighted in recent evidence as a critical contributor to trust among teachers, parents, and students (e.g., Bryk and Schneider 2003). This evidence suggests that principals engender trust with and among staff and with both parents and students when they:

- Recognize and acknowledge the vulnerabilities of their staff.
- Listen to the personal needs of staff members and assist as much as possible to reconcile those needs with a clear vision for the school.
- Create a space for parents in the school and demonstrate to parents that they (principal) are reliable, open, and scrupulously honest in their interactions.
- Buffer teachers from unreasonable demands from the policy environment or from the parents and the wider community.
- Behave toward teachers in a friendly, supportive, and open manner.
- Set high standards for students and then follow through with support for teachers.

Collective Teacher Efficacy

Also a teacher-level mediator, collective teacher efficacy (CTE) is the level of confidence a group of teachers feels about its ability to organize and implement whatever educational initiatives are required for students to achieve high standards of achievement. The effects of efficacy or collective confidence on performance are indirect through the persistence it engenders in the face of initial failure and the opportunities it creates for a confident group to learn its way forward (rather than giving up).

In highly efficacious schools, evidence suggests that teachers accept responsibility for their students' learning. Learning difficulties are not assumed to be an inevitable by-product of low SES, lack of ability, or family background. CTE creates high expectations for students as well as the collectively confident teachers. Evidence suggests that high levels of CTE encourage teachers to set challenging benchmarks for themselves, engage in high levels of planning and organization, and devote more classroom time to academic learning. High CTE teachers are more likely to engage in activity-based learning, student-centered learning, and interactive instruction. Among other exemplary practices, high CTE is associated with teachers adopting a humanistic approach to student management, testing new instructional methods to meet the learning needs of their students, providing extra help to students who have difficulty, displaying persistence and resiliency in such cases, rewarding students for their achievements, believing that their students can reach high academic goals, displaying more enthusiasm for teaching, committing to community partnerships, and having more ownership in school decisions.

While the total number of well-designed studies inquiring about CTE effects on students is still modest (about eight studies), their results are both consistent and impressive. This relatively recent corpus of research demonstrates a significant positive relationship between collective teacher efficacy and achievement by students in such areas of the curriculum as reading, math, and writing. Furthermore, and perhaps more surprising, several of these studies have found that the effects on achievement of CTE exceed the effects of students SES (e.g., Goddard, Hoy, & Woolfolk Hoy 2000) which, as we have already indicated, typically explains by far the largest proportion of achievement variation across schools. High CTE schools also are associated with lower suspension and dropout rates as well as greater school orderliness (Tschannen-Moran and Barr 2004).

There are two sources of insight about how leaders might improve the collective efficacy of their teaching colleagues. One source is the theoretical work of Albert Bandura, clearly the major figure in thinking about CTE. His work, now widely supported empirically, identified a number of conditions which influence the collective efficacy of a group: opportunities to master the skills needed to do whatever the job entails, vicarious experiences of others performing the job well, and beliefs about how supportive is the setting in which one is working. Leaders have the potential to influence all of these conditions, for example, by:

- Sponsoring meaningful professional development.
- Encouraging their staffs to network with others facing similar challenges in order to learn from their experiences.
- Structuring their schools to allow for collaborative work among staff.

A second source of insight about how leaders might improve the collective efficacy of their teaching colleagues is the small number of studies that have inquired about the leadership practices which improve CTE. For the most part, these have been studies of transformational leadership practices on the part of principals. Evidence from these studies demonstrates significant positive effects on CTE when principals:

- Clarify goals by, for example, identifying new opportunities for the school, developing (often collaboratively), articulating and inspiring others with a vision of the future, promoting cooperation and collaboration among staff toward common goals.
- Offer individualized support by, for example, showing respect for individual members of the staff, demonstrating concern about their personal feelings and needs, maintaining an open door policy, and valuing staff opinions.
- Provide appropriate models of both desired practices and appropriate values ("walking the talk").

The three mediators of leadership effects on student learning discussed in this section, like a large proportion of the larger set identified by Leithwood and Levin (2005) have two qualities in common worth further attention. They are properties of the group and they are "soft"—sociopsychological states rather than bricks and mortar—money, contracts, or teaching materials. Both of these qualities make them quintessentially suitable for the attention of school-level leaders and their school improvement efforts. Those leaders physically in the school can act in ways that are more sensitive to the underlying beliefs, values, and emotions from which these school conditions spring. Furthermore, there is little dependence on resources controlled largely outside the school in order to nurture the development of conditions.

Leaders need to know, in sum:

- Which of a wide array of potential mediators should be a priority for their attention and effort.
- What the status is of each of these mediators in their schools.
- What they can do to improve the condition of each of these high priority mediators in their schools.

Challenge Five: Overcoming Common Errors in Human Judgment

The challenge described in this section is one that school leaders frequently encounter. One manifestation of this challenge is the tendency for some teachers to quickly dismiss the results of large-scale assessment information about their students' performance when it deviates significantly from the results of their own assessments. This makes it very difficult for school leaders to engage their staffs in serious deliberations about how to interpret and use large-scale assessment information in their schools. The information, from their teachers' perspective, is largely invalid. Our analysis of the challenge, of which this is an instance, draws heavily on a strand of psychological research with roots that can be traced back more than 150 years, but one which acquired significant empirical traction beginning in the 1970s thanks to the efforts of such scholars as Nisbett and Ross (1980) and Kahneman et al. (1982).

This challenge, to be clear at the outset, is unique neither to administrators and teachers nor to the processes they use to make sense of large-scale assessment results, in particular. It is a pervasive challenge confronted whenever judgements are being made under less than "certain conditions," which is very often. Necessarily inferential in nature, judgements under uncertainty often fall prey to the kinds of errors that scientists have worked hard to develop methods for avoiding in their own work. These methods, suitably adapted to everyday judgment and choice, serve as the standards for determining whether or not a person or group, acting as "intuitive scientists" (Nisbett and Ross 1980), has committed errors in judgment.

Over a period of about 10 years, from the mid 1980s to the mid 1990s, my colleagues and I studied the judgment and problem-solving processes of educational administrators (principals and superintendents) in an effort to identify differences between expert and typical administrators (e.g., Leithwood and Steinbach 1995). Some of those studies were framed by concepts drawn from research on human judgment carried out by the authors cited above, among others. These studies were aimed at determining the extent to which differences in the expertise of school administrators could be accounted for by the types and incidence of errors made in their judgment processes. Our evidence suggested that such differences did account for significant variation in administrators' expertise and illustrated the nature of this variation.

Table 2.2 Errors in human judgment applied to evidence about student achievement

Types of errors	Example in context of large-scale assessment results
1. Overweighting vividness in interpreting the problem	Discount statistical information about student achievement in favor of own case-based impressions
2. Generalizing from a small or biased sample	Allow experiences with small group of students to overwhelm judgements about what to do with other students
3. Failure to see that a situation is unique or different from others in the past	Discount the need for changes in instruction due to changes in student cohort
4. Failure to determine actual causes of problem. (The previous section provided an extended treatment of this error)	Move directly from report of student achievement to instructional strategies without considering causes of failure
5. Failure to modify a single approach or strategy in light of situational features	Continue "tried and true" instructional strategies in face of significant changes in student population
6. Use of theories and schemas that do not accurately represent reality	Schools cannot compensate for challenges to student learning caused by family background

The left column of Table 2.2 lists those cognitive errors identified by Nisbett and Ross (1980) which were included in one of our studies of leadership expertise (Stager and Leithwood 1989). The right column provides examples of how each category of error, on the part of either teachers or school leaders, might manifest itself in the context of responding to large-scale assessment results. Six types of errors are included in the table: overweighting vividness in setting priorities or interpreting the problem; generalizing from a small or biased sample; and four errors that have in common the overuse or misuse of one's existing theories, knowledge structures, or schemas. Error 4 was the focus of Challenge Two (above) so it is not discussed any further here.

Overweighting Vividness in Interpreting the Problem

This error consists of the tendency to ignore very robust data about student achievement in statistical form in favor of much more immediate, vivid, and multidimensional impressions. Typically, these will be impressions gleaned from observing students in one's own class, in the case of teachers, or in the last class you visited, in the case of principals. Such data are vivid because they have real people attached to them. This error might lead to an assertion such as "these tests are not sensitive enough to capture what our students really know."

School leaders and teachers repeatedly demonstrating this error in response to large-scale assessment results will benefit from opportunities to develop a better understanding of basic concepts (e.g., standard deviation, scale reliability) used in summarizing data in statistical form. The goal, in this case, is to create vividness where it does not exist by giving people the tools to "see it" in a form different from what they have been are used to. Many school leaders will benefit from these same opportunities.

Generalizing from a Small or Biased Sample

This error will sometimes become evident in the tendency for teachers to generalize to the whole class or the whole school the performance of those students with whom they are most familiar. This error does not arise from vividness but from insensitivity to the wide variation found in larger populations on most matters. School leaders will need to provide targeted opportunities for teachers prone to improve their understanding of the pitfalls of generalizing from small samples.

Failure to See That a Situation Is Unique or Different from Others in the Past

Teachers and school leaders who are expert at what they do are less prone to this error than are their less expert counterparts. Nevertheless, both the fast-paced nature of the world they work in and the automaticity that is part of becoming expert in one's field leave even experts at risk of this error. Both fast pace and automaticity press teachers and administrators confronted with new problems to search for similarities with problems they have experienced and solved in the past. The detection of similarities will trigger well-rehearsed solutions, thereby reducing the cognitive demand required for a response.

This error might easily creep into the responses of principals and teachers examining the aggregated mean results of their students' performance on this year's large-scale assessment. Should such aggregated results mirror, or be very similar to, the aggregated mean results of last year's assessment, it would be easy for them to conclude that nothing has changed and to give the data no further thought. In fact, the aggregated results might well mask significant differences from last year on the part of some groups of students. Results for the small cohort of ESL students, for example, might have fallen dramatically while the results for all other students has crept forward just enough to keep the average level of achievement unchanged. Too bad for the ESL students!

This is a remarkably difficult error for school leaders and their staffs to avoid for the reasons already mentioned (automaticity and fast-paced context). It requires at least constant vigilance on the part of leaders and a willingness to ask questions of staff which will, at the time, seem only to slow down decision making and create more work. This likely demands an exceptionally reflective disposition on the part of school leaders and a willingness to foster such a disposition among teachers.

Failure to Modify a Single Approach or Strategy in Light of Situational Features

This error may occur even when people acknowledge unique features of the problem they are facing in comparison with problems they have addressed in the past. This error entails acting on the assumption that one's typical or previously used solutions will be suitable or powerful enough to succeed anyway. So school staffs may acknowledge, for example, that the new curriculum guidelines issued by the province place much greater emphasis on students' own knowledge construction while, nevertheless, also maintain that their current forms of instruction, premised on a very different model of how students learn, will still be suitable.

Returning to an earlier example, suppose the same school staff faced with unchanged year-over-year aggregated mean large-scale assessment results were eventually persuaded to disaggregate the data and found the problem with their ESL students. If they then decided to continue with their "tried and true" instructional strategies, they would be guilty of this error. To reduce the incidence of this error, school leaders will need to ask their colleagues uncomfortable questions about the justification for continuing practices that seem unlikely to be productive in changed circumstances.

Making Use of Theories or Schemas That Do Not Accurately Represent Reality

This error is very well illustrated by the results of a recent analysis conducted in Australia (Mulford et al. 2007). The study compared a large sample of elementary and secondary school principals' estimates of how well their students were achieving in literacy and numeracy with students' actual scores on state tests; student success was classified, for this purpose, as low, medium, or high. Results of this study pointed toward a strong tendency for principals of schools whose actual student scores were low and medium to significantly overestimate the success of their students (74% overestimated for primary and 71% for secondary), in some cases by two levels (16% for primary and 30% for secondary).

A larger proportion of principals of schools whose students were actually high achieving (86% primary and 63% secondary) estimated such achievement accurately. These results, however, cannot be interpreted to suggest that principals of higher achieving schools are more accurate in their assessments, only that the tendency of principals to be "optimistic" about their school's success had a greater chance of reflecting reality in high performing schools. The errors in principals' estimates of their students' success, it should be noted, occurred in a policy context which makes test scores on state exams widely available to schools and encourages their use!

The cognitive error which surfaced in this study also reflects evidence summarized by Nisbett and Ross (1980) indicating that peoples' beliefs and behaviors have a tendency to become aligned over time even when they start off being very different. Arguably, principals are under pressure from many sources to be cheerleaders for their schools. After months or years of defending or justifying the quality of their school programs, they may well begin to believe the justification—even if that was not the case at the outset.

Avoiding this error likely requires a high level of metacognitive control on the part of the principal. In its absence, principals will lose the critical edge they need to continue moving their school forward. This is one explanation for the widespread belief in many organizations that leaders should not remain in the same position beyond 6 or 7 years (Gabarro 1987).

Conclusion

The overwhelming motivation for developing large-scale assessment programs has been the desire to hold schools more publicly accountable. I accept the need for some form of external accountability for schools and have not attempted, myself, to offer a better solution than large-scale assessment programs. My concern in this chapter has been with the proliferation of purposes for large-scale assessments that have occurred over time, and in particular, the claim that the results of such assessments are powerful sources of insight for those in schools, such as school leaders, aiming to improve student performance. The skeptics among us might be inclined to the view that this proliferation of purposes is little more than an attempt to justify the expenditure of sizeable public monies and the enormous opportunity costs associated with the student, teacher, and administrator time these assessments consume.

Most public opinion polls, nevertheless, suggest very high and stable levels of support for large-scale student testing programs (e.g., Livingstone et al. 2001, 2003). These same polls also indicate that very few respondents have much understanding of what such testing entails, the technical limitations of what they produce, and the collateral outcomes which often result. With such blind support, there is little incentive among assessment advocates to do anything but continue. Like it or not, school leaders will have to deal with the results of large-scale assessments, "warts and all," in the foreseeable future. So my purpose in this chapter has been to unpack a series of challenges faced by improvement-oriented school leaders when confronted with the results of large-scale assessments for their schools.

Adopting a cognitive psychological perspective, I framed the student performance improvement task as an ill structured problem—ill structured because of the uncertain nature of the "solution paths" between a current state (the current level of student performance in one' school) and a goal state (the desired, and no doubt, higher, level of such performance). With this perspective on the improvement task of school leaders, the chapter asked in what ways the results of a typical large-scale assessment could be helpful. It is reasonable to expect that assessment results, as a minimum, would help school leaders and their staffs understand the current state of student learning in their schools. But large-scale assessments actually have very limited potential for this very important purpose because of the narrow set of student outcomes they measure, their unknown reliability at the school and classroom levels, difficulties in using such data for tracking changes in student performance over time, and the large time lags between collecting and reporting results. These unhelpful features of large-scale assessment programs, furthermore, are as serious a compromise to their value in determining whether a desired state of student performance has been reached as they are in determining the current state.

Beyond pointing out these limitations of large-scale assessments, the chapter described several other critical challenges that school leaders face in their efforts to improve their students' performance. Two of these challenges can only be addressed with information that most teachers and school leaders acquire through informal and almost always unsystematic means, at best. This is information about the causes of students' current levels of achievement and the status of those conditions in the school which nurture and support its improvement efforts. We need to move past the view that information about the status of students' achievement is the only information needed for school improvement purposes. School leaders desperately need robust, systematically-collected information about those other features of their schools that account for student achievement if their success is to approximate the aspirations of our current policies.

This chapter, in sum, has argued that most large-scale assessment results are of quite limited practical use to school leaders and their teacher colleagues in their efforts to improve their students' performance. Such results, for example, might just as easily misrepresent, as accurately capture, current levels of achievement in modest-sized schools and in virtually all classrooms. Evidence about multi-year achievement patterns across Ontario schools indicated just how uneven is the trajectory of improvement in school performance when Ontario's large-scale assessment program is the yardstick for measuring that performance. Admittedly, schools occasionally do face circumstances with the documented potential to quickly and dramatically alter the quality of their students' educational experiences. Principal succession is one of these circumstances (MacMillan 2000); a high rate of teacher turnover with immediate effects on teacher quality is another. But these are occasional rather than frequent circumstances faced by schools and certainly do not occur sufficiently often in most schools to account for the bumpy trajectories of achievement found in the Ontario data. Changes in the quality of education provided to students by most schools, most of the time, are better described as gradual and highly incremental. Ironically, reformists wring their hands about such slow and incremental change in schools, on the one hand, yet are willing to accept large-scale assessment evidence of dramatic short term achievement increases and decreases at face value, on the other.

In the face of their large-scale assessment results, school leaders often feel like technically naïve, statistics virgins, struggling to unlock some powerful new insight buried in the numbers, if only they understood them. The sense of guilt produced by these feelings of inadequacy is not warranted, however. Similarly, there is a substantial literature that is quite critical of teacher assessment practices while arguing for an increase, on teachers' parts, in the skills and understandings more closely associated with the technology of large-scale assessments. This seems like a classic case of "the pot calling the kettle black." Certainly many teachers could benefit from additional understandings about test and measurement concepts. But this would not be of a different magnitude than the additional understandings needed by those promoting the use of large-scale assessment results for school improvement purposes. As it stands now, teachers have no reason to apologize for their own assessment efforts in the face of the huge challenges still to be addressed by the designs of most large-scale assessment programs (see the 2005 annual yearbook of the *National Society of Education* for much more on this). Flaw for flaw—and I claim only impressionistic evidence here—most teachers' assessments seem likely to provide better clues for instructional improvement than do most large-scale assessment results. Those of us who are members of the research and policy communities should be much more forthright with teachers and administrators about the limitations of large-scale assessments for school improvement purposes.

References

Alig-Mielcarek, J. M. (2003). *A model of school success: Instructional leadership, academic press, and student achievement.* Unpublished doctoral dissertation, Ohio State University, Columbus, OH.

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in preK-12 education. *Educational Researcher, 29*(8), 24–25.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability (103rd yearbook of the national society for the study of education)* (pp. 20–50). Chicago, IL: University of Chicago Press.

Bryk, A. S., & Schneider, B. (2003). Trust in schools: A core resource for school reform. *Educational Leadership, 60*(6), 40–44.

Creemers, B. (2007). Educational effectiveness and improvement: The development of the field in mainland Europe. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 223–241). Dordrecht, the Netherlands: Springer.

Creemers, B. P. M., & Reetzig, G. J. (1996). School level conditions affecting the effectiveness of instruction. *School Effectiveness and School Improvement, 7*(3), 197–228.

Dirks, K. T., & Ferrin, D. L. (2002). Trust in leadership: Meta-analytic findings and implications for research and practice. *Journal of Applied Psychology, 87*(4), 611–628.

Earl, L. M. (2001). Data, data everywhere (and we don't know what to do): Using data for wise decisions in schools. In P. de Broucker & A. Sweetman (Eds.), *Towards evidence-based policy for Canadian education* (pp. 39–51). Kingston, ON: John Deutsch Institute for the Study of Economic Policy, Queen's University.

Fredericksen, N. (1984). Implications of cognitive theory for instruction in problem solving. *Review of Educational Research, 54*(3), 363–407.

Gabarro, J. J. (1987). *The dynamics of taking charge.* Boston, MA: Harvard Business School Press.

Gagné, E. D. (1985). *The cognitive psychology of school learning.* Boston, MA: Little, Brown and Co.

Goddard, R. D. (2003). Relational networks, social trust, and norms: A social capital perspective on students' chance of academic success. *Educational Evaluation and Policy Analysis, 25*(1), 59–74.

Goddard, R. D., & Goddard, Y. L. (2001). A multilevel analysis of the relationship between teacher and collective efficacy in urban schools. *Teaching and Teacher Education, 17*(7), 807–818.

Goddard, R. D., Hoy, W. K., & Woolfolk Hoy, A. (2000). Collective teacher efficacy: Its meaning, measure and impact on student achievement. *American Educational Research Journal, 37*(2), 479–507.

Goddard, R. D., Sweetland, S. R., & Hoy, W. K. (2000). Academic emphasis of urban elementary schools and student achievement in reading and mathematics: A multi-level analysis. *Educational Administration Quarterly, 36*(5), 683–702.

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

Heubert, J., & Hauser, R. (Eds.). (1999). *High stake tests: Testing for tracking, promotion and graduation.* Washington, DC: National Academic Press.

House, R., Hanges, P., Javidan, M., Dorfman, P., & Gupta, V. (2004). *Culture, leadership and organizations: The Globe study of 62 societies.* Thousand Oaks, CA: Sage.

Jacob, J. A. (2004). *A study of school climate and enabling bureaucracy in select New York City public elementary schools.* Unpublished doctoral dissertation, University of Utah, Salt Lake City, UT.

Jurewicz, M. M. (2004). *Organizational citizenship behaviors of middle school teachers: A study of their relationship to school climate and student achievement.* Unpublished doctoral dissertation, College of William and Mary, Williamsburg, VA.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Lee, V. E., & Croninger, R. G. (1994). The relative importance of home and school in the development of literacy skills for middle-grade students. *American Journal of Education, 102*(3), 286–329.

Leithwood, K., Day, C., Sammons, P., Harris, A., & Hopkins, D. (2006). *Successful school leadership: What it is and how it influences pupil learning.* London, UK: DfES. Available at http://www.dfes.gov.uk/research/data/uploadfiles/RR800.pdf.

Leithwood, K., & Levin, B. (2005). *Assessing school leader and leadership programme effects on pupil learning* (RR662). Department for Education and Skills (DfES).

Leithwood, K., & Riehl, C. (2005). What we know about successful school leadership. In W. Firestone & C. Riehl (Eds.), *A new agenda: Directions for research on educational leadership* (pp. 22–47). New York, NY: Teachers College Press.

Leithwood, K., Seashore Louis, K., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning: A review of research for the Learning from Leadership Project.* New York, NY: The Wallace Foundation.

Leithwood, K., & Steinbach, R. (1995). *Expert problem solving processes: Evidence from principals and superintendents.* Albany, NY: SUNY Press.

Linn, R. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3–13.

Livingstone, D. W., Hart, D., & Davie, L. E. (2001). *Public attitudes towards education in Ontario 2000: The 13th OISE/UT Survey.* Toronto, ON: Ontario Institute for Studies in Education, University of Toronto.

Livingstone, D. W., Hart, D., & Davie, L. E. (2003). *Public attitudes towards education in Ontario 2002: The 14th OISE/UT Survey.* Toronto, ON: Ontario Institute for Studies in Education, University of Toronto.

Macmillan, R. (2000). Leadership succession, cultures of teaching and educational change. In N. Bascia & A. Hargreaves (Eds.), *The sharp edge of educational change: Teaching, leading and the realities of reform* (pp. 52–71). London, UK: Routledge/Falmer.

McDonnell, L. M. (2005). Assessment and accountability from the policy maker's perspective. In J. Herman & E. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement (104th Yearbook of the National Society for the Study of Education)* (pp. 35–54). Malden, MA: Blackwell.

Mulford, B., Kendall, D., Edmunds, B., Kendall, L., Ewington, J., & Silins, H. (2007). Successful school leadership: What is it and who decides? *Australian Journal of Education, 51*(3), 228–246.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Pitner, N. (1988). The study of administrator effects and effectiveness. In N. Boyan (Ed.), *Handbook of research on educational administration* (pp. 99–122). New York, NY: Longman.

Sackney, L. (2007). History of the school effectiveness and improvement movement in Canada over the past 25 years. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 167–182). Dordrecht, the Netherlands: Springer.

Silins, H., & Mulford, B. (2007). Leadership and school effectiveness and improvement. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 635–658). Dordrecht, the Netherlands: Springer.

Stager, M., & Leithwood, K. (1989). Cognitive flexibility in principals' problem solving. *Alberta Journal of Educational Research, 35*(3), 217–236.

Teddlie, C., & Stringfield, S. (2007). A history of school effectiveness and improvement research in the USA focusing on the past quarter century. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 131–166). Dordrecht, the Netherlands: Springer.

Timperley, H. S. (2005). Distributed leadership: Developing theory from practice. *Journal of Curriculum Studies, 37*(6), 395–420.

Tschannen-Moran, M., & Barr, M. (2004). Fostering student achievement: The relationship between collective teacher efficacy and student achievement. *Leadership and Policy in Schools, 3*(3), 189–209.

Walberg, H. (1984). Improving the productivity of America's schools. *Educational Leadership, 41*(8), 19–27.

Wilson, D. (2004). Assessment, accountability and the classroom: A community of judgment. In D. Wilson (Ed.), *Towards coherence between classroom assessment and accountability (103rd Yearbook of the National Society for the Study of Education)* (pp. 1–19). Chicago, IL: University of Chicago Press.

Wolfe, R., Childs, R., & Elgie, S. (2004). *Final report of the external evaluation of the EQAO's assessment process.* Toronto, ON: Ontario Institute for Studies in Education, University of Toronto.