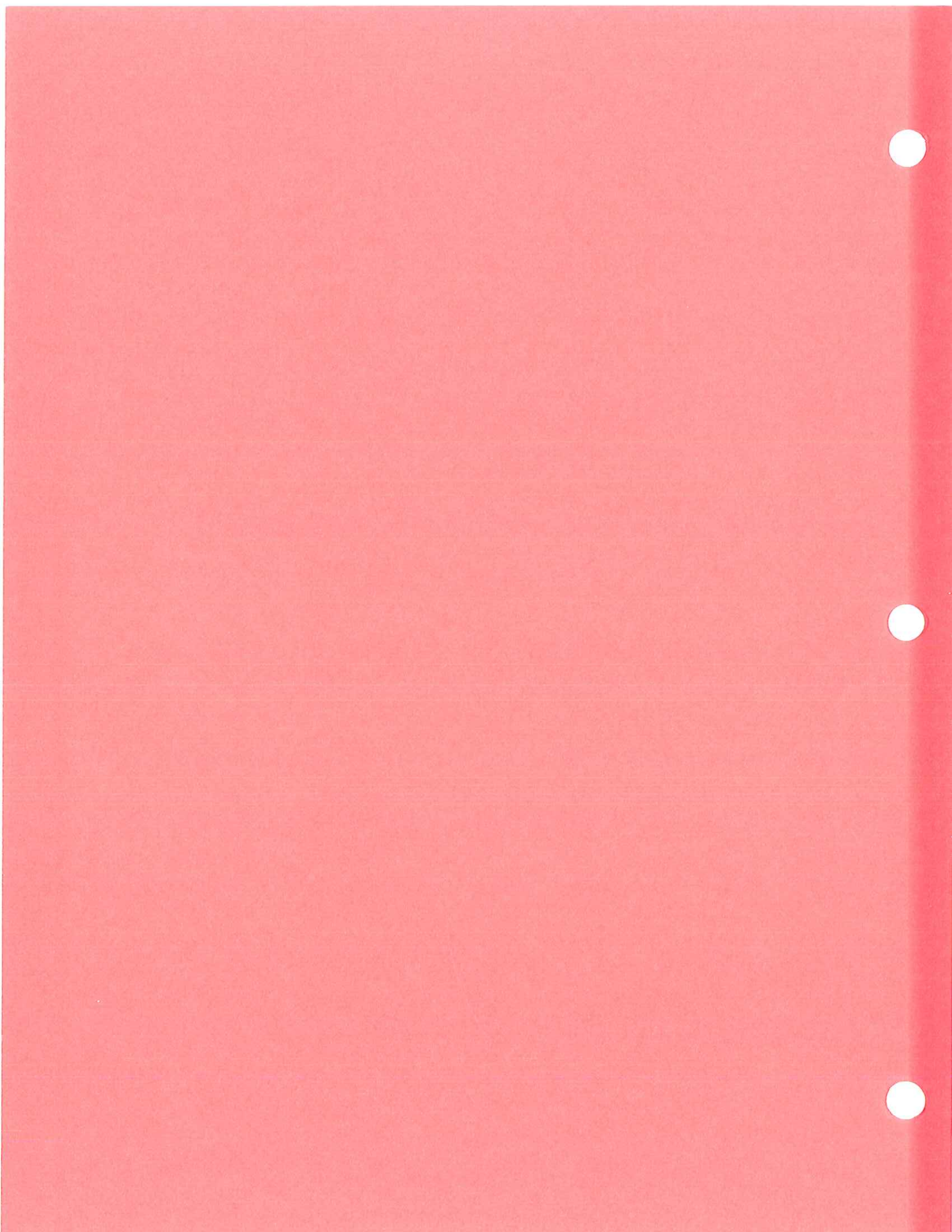# Douglas Reeves

Chairman and Founder
Center for Performance Assessment

Nominated by
Debbie Arato

Douglas Reeves
Chairman and Founder, Center for Performance Assessment

Dr. Douglas Reeves is chairman and founder of the Center for Performance Assessment, an international organization dedicated to improving student achievement and educational equity. Through its long-term relationships with school systems, the Center helps educators and school leaders to improve student achievement through practical and constructive approaches to standards, assessment, and accountability.

Dr. Reeves is a frequent keynote speaker in the U.S. and abroad for education, government, and business organizations and is a faculty member of leadership programs sponsored by the Harvard Graduate School of Education. The author of nineteen books and many articles, Doug is the author of the best-selling, *Making Standards Work* now in its third edition. Other recent titles include *Assessing Educational Leaders: Evaluating Performance for Improved Individual and Organizational Results* (Corwin Press, 2004), *Accountability for Learning: How Teachers and School Leaders Can Take Charge* (ASCD, 2004), *The Daily Disciplines of Leadership: How to Improve Student Achievement, Staff Morale, and Personal Organization* (Jossey Bass, 2002), *The Leader's Guide to Standards: A Blueprint for Educational Excellence and Equity* (Jossey-Bass, 2002), and *Reason to Write: Help Your Child Succeed in School and In Life Through Better Reasoning and Clear Communication* (Simon & Schuster, 2002). Doug has twice been selected for the Harvard Distinguished Author's Series and he recently won the Parent's Choice Award for his writing for children and parents. His writings have been translated into Chinese, Japanese, Hebrew, and Spanish.
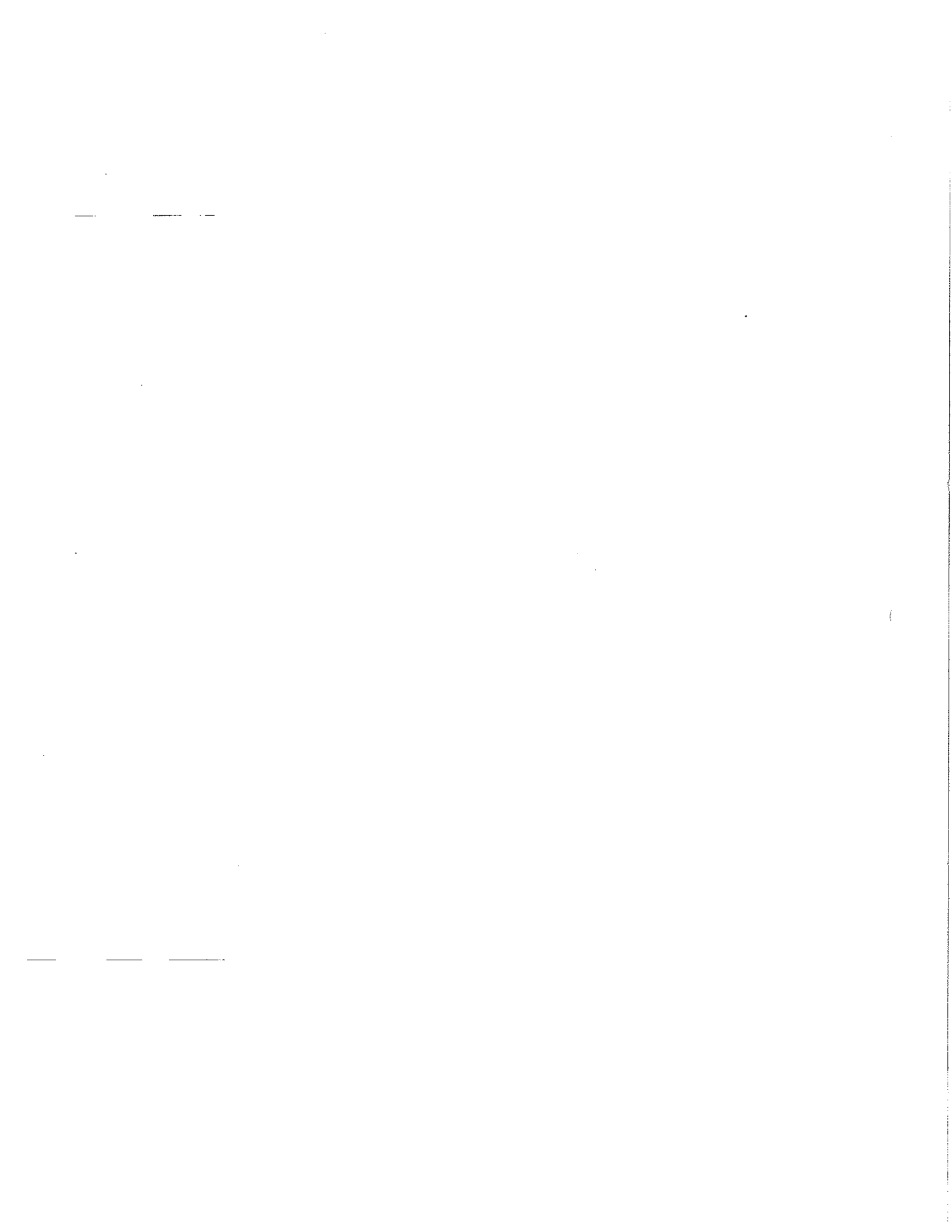
Beyond his work in large-scale assessment and research, Doug has devoted many years to classroom teaching with students ranging from elementary school to doctoral candidates. Doug's family includes four children ranging from elementary school through college, all of whom have attended public schools. His wife, Shelley Sackett, is an attorney, mediator, and school board member.

Submitted by: Debbie Arato

WORKS OF DOUGLAS REEVES, PH.D.

Douglas B. Reeves, Ph.D.

225 Derby Street, Suite 503
Salem, MA 01970
(978) 740-3001   Fax: (978) 740-3002
Email: dreeves@makingstandardswork.com

Positions:

Chairman, Center for Performance Assessment
Faculty member, Harvard Leadership and Policy Forum

What Leading Educational Writers Say About Dr. Reeves:

"There is not a person in the United States who better understands standards."
-Bill Habermehl, superintendent, Orange County, California Office of Education

"No writer in America has been more effective in presenting a persuasive rationale for the ethical imperative of articulating standards of learning for all students than Dr. Douglas Reeves"
-Richard DuFour, Author, On Common Ground

"At a time when school leadership is central to the success of students, Reeves offers a comprehensive means to measure the critical elements of effective school leadership."
-Vincent Ferrandino, Executive Director National Association of Elementary School Principals

"Research studies increasingly support the significant impact that effective leaders have on school performance. A skilled leader is an essential component in successful school reform. This book makes a compelling case for transforming leadership evaluation from perfunctory to a focus on performance, reflection, and results."
-Gerald N. Tirozzi, Executive Director, National Association of Secondary School Principals

"This book provides an answer to the abiding question that must be embedded in the thought process of all effective leaders: 'How can I add value to my life and the life of others?' It is a must read!"
-Marti Richardson, President, National Staff Development Council

"*The Leader's Guide to Standards* is an important resource for anyone who cares about getting the standards and accountability system right. This book gives sound guidance to the many stakeholders who must work together to ensure that our kids succeed. It is a must read."
-Anne Bryant, Executive Director, National School Boards Association

"*The Leader's Guide to Standards* prepares leaders to engage in informed, sustained conversation within the education community about the values, goals, and methods of standards-based education."
    -Dennis Sparks, Executive Director, National Staff Development Council

"This book brings together music of the cumulative wisdom gained and imparted by Doug in one engaging and compelling source. Educators at all levels will benefit for reading it."
    -Robert J. Marzano, Senior Scholar, Mid-continent Research for Education and
    Learning Associate Professor, Cardinal Stritch University

Books (partial list):

It's All About the Client – Consulting for Results, Advanced Learning Press, 2005.

On Common Ground- The Power of Professional Learning Communities, National
    Educational Service, 2005

101 More Questions and Answers about Standards, Assessment, and Accountability,
    Advanced Learning Press, 2004.

Accountability for learning: How teachers and school leaders can take charge, ASCD,
    2004. (Also translated into Chinese.)

*Assessing Educational Leaders: Evaluating performances for improved individual and
    organizational results,* Corwin, 2003.

Leader's guide to standards: A blueprint for educational equity and excellence, Jossey
    Bass, 2002.

*Daily disciplines of leadership: How to improve student achievement, staff motivation,
    and personal organization,* Jossey Bass, 2002.

*Reason to write: Help your child succeed in school and in life through better reasoning
    and clear communication,* Kaplan, 2002.

Reason to Write Student Handbook, Kaplan, 2002.

Holistic accountability: Serving students, schools, and community, Corwin
    Press, 2002.

*20-Minute learning connection: A practical guide for parents who want to
    help their children succeed in school,* Simon & Schuster, 2001.

*Crusade in the classroom: How George W. Bush's education reforms will
    affect your children, our schools,* Simon & Schuster, 2001.

101 questions and answers about standards, assessment, and accountability,
    Advanced Learning Press, 2000.

Accountability in action: A blueprint for learning organizations, Advanced
    Learning Press, 1999.

*Making standards work: How to implement standards-based performance
    assessments in the classroom, school, and district* (2nd ed.), Advanced
    Learning Press, 1996.

Articles (partial list):

"Constructive Alternative in a Destructive Debate", *Principal Leadership*, March 2005
"Accountability at the Crossroads", *Leadership*, December 2004
"The Case Against the Zero", *Phi Delta Kappan,* December 2004
"Evaluating Administrators," *Educational Leadership,* April 2004
"Leave me alone and let me teach," *School Administrator*, December 2001.
"If you hate standards, learn to love the bell curve," *Education Week*, June 2001.
"Standards make a difference" *NASSP Bulletin*, January 2001.
"Standards are not enough: Essential transformations for successful schools," *NASSP Bulletin*, December 2000.
"Three keys to professional development," *California Curriculum News Report*, February 2000.
"Finishing the race," *Thrust for Educational Leadership*, May/June 2000.
"Holding principals accountable," *School Administrator*, October 1998.
"Responding to the rhetoric of the radial right," *School Administrator*, March 1998.
"Practical performance assessment for busy teachers," *Learning Magazine*, January/February, 1998.
"Defending performance assessment without being defensive," *School Administrator*, June 1997.

Dr. Reeves has also been interviewed and cited in numerous sources, including *USA Today, Washington Post, International Herald-Tribune, Boston Globe, US News & World Report, Los Angeles Times, San Francisco Chronicle, Deseret News*, and numerous regional, local, and scholarly publications.

Presentations (partial list):

National Association of Elementary School Principals, 2005, Baltimore, leadership address
American Association of School Administrators, 2005, San Antonio, keynote address
National Association of Elementary School Principals, 2004, San Diego, leadership address
National School Board Association, 2004, Orlando, leadership address
Association of Supervision and Curriculum Development, 2004, New Orleans, leadership address
National Leadership Conference, 2002, San Diego, keynote address.
National School Board Association, 2002, New Orleans, leadership address.
California Superintendents Association, Education Writer Award Address, 2002.
Harvard Distinguished Author's Series, 2002, Cambridge, author remarks.
Effective Schools Conference, 2002, Phoenix, keynote address.
National Standards & Assessment Conference, 2002, Las Vegas, keynote address.
Northeast ASCD, 2001, Boston, featured presentation.
National Education Association, 2001, Atlanta, keynote address.
Harvard Leadership and Policy Forum, 2001, Cambridge, keynote address.
Numerous state, regional, national, and international keynote speeches.

Education:

B.A., University of Wyoming, *Phi Beta Kappa*, 1975.
M.A., University of Southern California, distinguished graduate, 1979.
Ph. D., University of Wyoming, concentration in research, measurement, and
statistics, 1994.

Teaching Experience:

Hillel Academy (elementary and middle school mathematics)
Rivers Barracks, Giessen, Germany (high school G.E.D. and advanced placement
American government and American history)
Chuzho Middle School, Anhui province, China (graduate teaching practice and
English as a second language)
University of Wyoming (high school debate, undergraduate political science,
graduate research, and graduate statistics)
Regis University (graduate leadership)
Colorado Christian University (undergraduate leadership and decision-making,
graduate statistics and research, graduate educational psychology)
Harvard University (Graduate School of Education Principal's Center, courses in
assessment, leadership, and educational policy)

Awards and Commendations:

Parents Choice Award (2002)
Education Writer of the Year (California Superintendent's Association, 2002)
Distinguished author (Harvard Graduate School of Education, 2002)
Meritorious Service Medal (Oak leaf cluster), United States Army Military Intelligence
Distinguished Graduate, University of Southern California
Highest Honors, University of Wyoming

Personal:

Married to Shelley Sackett. Four children.

19

CHAPTER

The 90/90/90 Schools: A Case Study

Research conducted at the Center for Performance Assessment on the "90/90/90 Schools" has been particularly instructive in the evaluation of the use of standards and assessment. The research includes four years of test data (1995 through 1998) with students in a variety of school settings, from elementary through high school. Our analysis considered data from more than 130,000 students in 228 buildings. The school locations included inner-city urban schools, suburban schools, and rural schools. The student populations ranged from schools whose populations were overwhelmingly poor and/or minority to schools that were largely Anglo and/or economically advantaged.

One reason that the research in these schools was so productive is that the districts maintained careful records on actual instructional practices and strategies. This allows researchers to investigate associations between instructional strategies and academic achievement results. It is important to acknowledge, however, that these results are only associative in nature. We make no claim that a single instructional intervention can be said to "cause" a particular achievement result. What we can say with a high degree of confidence, however, is that there are some consistent associations between some classroom strategies (for example, performance assessments that require writing) and student achievement in a wide variety of tests and subjects. One final note: We make absolutely no claim that the schools in the study were the beneficiaries of any proprietary "program" or "model" of instruction.

The research literature in every field from pharmaceuticals to education contains too many "studies" that purport to show the effectiveness of treatments that the authors of the research have used. Our role in this investigation is that of journalist and researcher, not of architect of any program or intervention. Hence, we do not claim any credit for improved academic achievement that is rightfully due to the students, teachers, and administrators in the study.

Characteristics of 90/90/90 Schools

The 90/90/90 Schools have the following characteristics:

- More than 90 percent of the students are eligible for free and reduced lunch, a commonly used surrogate for low-income families.

- More than 90 percent of the students are from ethnic minorities.

- More than 90 percent of the students met or achieved high academic standards, according to independently conducted tests of academic achievement.

The educational practices in these schools are worthy of notice for several reasons. First, many people assume that there is an inextricable relationship between poverty, ethnicity, and academic achievement. The graph in Figure 19.1 expresses the commonly held belief that poverty and ethnic minority enrollment are inextricably linked to lower levels of student achievement.

Figure 19.1

Common Assumptions About the Relationship Between Poverty, Minority Enrollment, and Student Achievement

$r = -.6$ to $.9$

Percentage Proficient

Percent Free and Reduced Lunch and Minority Enrollment
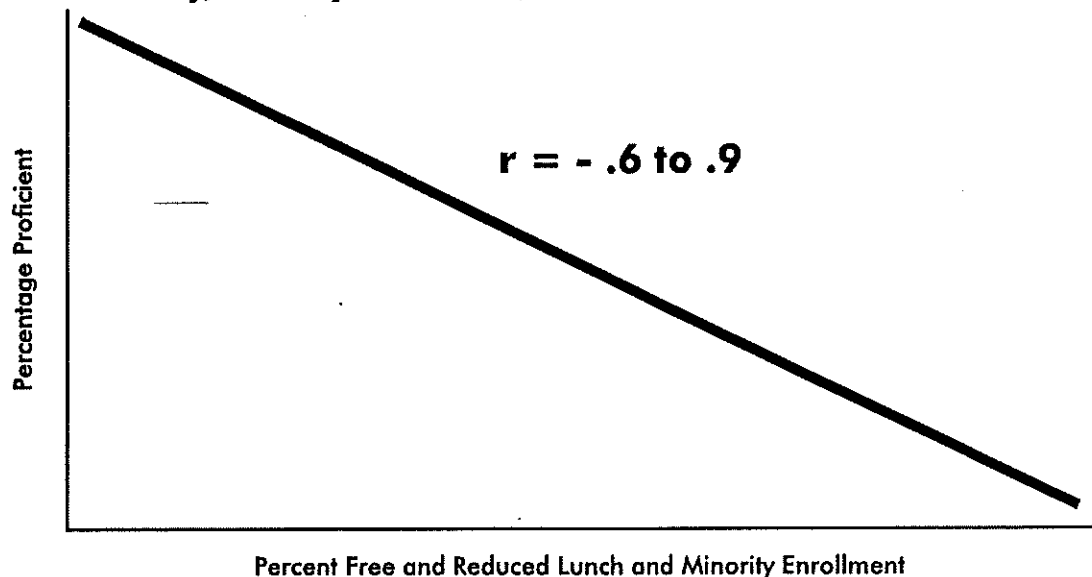
In this chart, the prevailing hypothesis leaves no room for students in the upper right-hand corner of the graph—that is, schools that have high academic achievement coincident with high poverty and high minority enrollments. This is consistent with national observations dating back to the 1960s in which demographic characteristics were regarded as the dominant variables influencing student achievement. In fact, the actual data from the December 1998 Comprehensive Accountability Report of the Milwaukee Public Schools shows a different story. In individual schools, there are striking numbers of students who are

poor and who are members of ethnic minorities who also academically proficient. Throughout the entire system of more than 100,000 students, the relationship between poverty and student achievement is not the postulated -.6 to -.9, but rather a -.2. While the impact of poverty clearly has not been eliminated, the prevailing hypothesis that poverty and ethnic minority status are invariably linked to low student achievement does not conform to the data.

Common Characteristics of High Achievement Schools

Our research on the 90/90/90 Schools included both site visits and analyses of accountability data. The site visits allowed us to conduct a categorical analysis of instructional practices. In the same manner that the authors of *In Search Of Excellence* (Peters and Waterman, 1982) identified the common practices of excellent organizations, we sought to identify the extent to which there was a common set of behaviors exhibited by the leaders and teachers in schools with high achievement, high minority enrollment, and high poverty levels. As a result, we found five characteristics that were common to all 90/90/90 Schools. These characteristics were:

- A focus on academic achievement
- Clear curriculum choices
- Frequent assessment of student progress and multiple opportunities for improvement
- An emphasis on nonfiction writing
- Collaborative scoring of student work

Focus on Academic Achievement

After visiting all of the 90/90/90 Schools, we noticed profound differences between the assessment and instructional practices of these schools and those of low-achieving schools. First and most importantly, the 90/90/90 Schools had a laser-like focus on student achievement. The most casual observer could not walk down a hallway without seeing charts, graphs, and tables that displayed student achievement information, as well as data about the continuous improvement students had made. The data were on display not only in principals' offices, but also throughout the schools. In addition, we saw school trophy cases full of exemplary academic work, including clear, concise essays, wonderful science projects, terrific social studies papers, and outstanding mathematics papers. In short, the 90/90/90 Schools made it clear to the most casual observer that academic performance was highly prized.

The focus on achievement in these schools included a particular emphasis on improvement. The comprehensive accountability system in use by these schools forced every school to identify five areas in which they measured improvement. Although the school could choose the goal from a menu, the common requirement was to focus on a few indicators of improvement in contrast to the typical school improvement plan that contains a large number of unfocused efforts to improve. The focus on improvement is especially important in an environment where many students come to school with academic skills that are substantially below grade level. The consistent message of charts showing weekly improvement from the fall through the spring was, "It's not how you start here that matters, but how you finish." Improvements of more than one grade level in a single year were common, and teachers and administrators paid particular attention to students whose deficiencies in reading and writing would have a profound impact on their success in other subjects. Some students spent as many as three hours per day in literacy interventions designed to get students to desired achievement levels. There did not appear to be any consistency with regard to the intervention programs in use by these schools. Some used Success for All, others used Reading Recovery, while others used the Efficacy Model. Others had no specified program at all, but consistently applied focused intervention for students in need using their own teaching staff.

Curriculum Choices

Such a focus on achievement inevitably leads to curriculum choices, spending more time on the core subjects of reading, writing, and mathematics and less time on other subjects. It is possible, for example, that many of the teachers in these schools did not "cover the curriculum" in the strict sense of checking off objectives from a wide variety of curricular areas. They chose—wisely, we believe—to emphasize the core skills of reading, writing, and mathematics in order to improve student opportunities for success in a wide variety of other academic endeavors later. It is interesting to note parenthetically that, despite their disproportionate emphasis on language arts and mathematics, these schools also significantly out-performed their peer schools on science tests as well. This makes an important point that eludes those who remain committed to a "coverage" model: tests of science, social studies, study skills, and virtually every other subject area are, in fact, tests of reading and writing.

Frequent Assessment of Student Progress with Multiple Opportunities for Improvement

Many of the high-poverty schools included students whose skills were significantly below grade level in academic achievement as they entered the school. The consistent message of the 90/90/90 Schools is that the penalty for poor performance is not a low grade, followed by a forced march to the next unit. Rather, student performance that is less than proficient is followed by multiple opportunities to improve performance. Most of these schools conducted weekly assessments of student progress. It is important to note that these assessments were not district or state tests, but were assessments constructed and administered by classroom

teachers. The consequence of students performing badly was not an admonishment to "Wait until next year" but rather the promise that "You can do better next week."

A frequent challenge to this practice is that students should learn to "get it right the first time." The flaw in such a statement is the implied assumption that the traditional "one-shot" assessment is successful in leading students to "get it right the first time." In fact, when students know that there are no additional opportunities to succeed, they frequently take teacher feedback on their performance and stuff it into desks, back packs, and wastebaskets. Students in this scenario are happy with a "D" and unmotivated by an "F." After all, there is nothing that they can do about deficient performance anyway. In a classroom assessment scenario in which there are multiple opportunities to improve, however, the consequence for poor performance is not a bad grade and discouragement, but more work, improved performance, and respect for teacher feedback. In this respect, the use of teacher evaluation based on assessment scoring guides looked much more like active coaching after which improvement was required, and much less like final evaluation from which there was no reprieve.

Written Responses in Performance Assessments

By far the most common characteristic of the 90/90/90 Schools was their emphasis on requiring written responses in performance assessments. While many schools with similar demographic characteristics employed frequent assessment techniques, many of the less successful schools chose to emphasize oral student responses rather than written responses. The use of written responses appears to help teachers obtain better diagnostic information about students, and certainly helps students demonstrate the thinking process that they employed to find a correct (or even an incorrect) response to an academic challenge. Only with a written response from students can teachers create the strategies necessary to improve performance for both teacher and learner.

In virtually every school we have evaluated, student scores on creative writing are significantly higher than informative and narrative writing scores. As a result, teachers in the successful 90/90/90 Schools placed a very high emphasis on informative writing. They typically used a single scoring rubric to evaluate student writing and applied this scoring guide to every piece of written work. Whether the student was writing a book report, lab report, social studies report, analysis of a sporting event, description of a piece of music, or a comparison of artists, the message was the same: this is the standard for good writing, and there are no compromises on these expectations for quality.

The benefits of such an emphasis on writing appear to be two-fold. First, students process information in a much clearer way when they are required to write an answer. They "write to think" and, thus, gain the opportunity to clarify their own thought processes. Second, teachers have the opportunity to gain rich and complex diagnostic information about why students respond to an academic challenge the way that they do. In contrast to the binary feedback (right/wrong) provided by most

assessments and worksheets, the use of performance assessments that require written responses allows the teacher to diagnose obstacles to student learning. By assessing student writing, teachers can discern whether the challenges faced by a student are the result of vocabulary issues, misunderstood directions, reasoning errors, or a host of other causes that are rarely revealed by typical tests.

The association between writing and performance in other academic disciplines was striking, and this gets to the heart of the curriculum choices that teachers must make. At the elementary level, for example, teachers were faced with a formidable set of curriculum standards in both science and writing. Many of the most successful schools reported that they had to sacrifice time allocated to every other curriculum area except reading, writing, and mathematics. Nevertheless, more than 80 percent of the 135 elementary schools in the study improved in science scores in 1998, compared to 1997. The Pearson correlation between writing improvement and science improvement is striking: .74—a large correlation in virtually any area of social science research. This correlation took place without any changes in the science curriculum and few apparent modifications in teaching methods. I would offer the same caution as provided earlier in the chapter that correlation is not causation. Nevertheless, when two variables appear to behave in such a similar way, it is difficult to escape the conclusion that an emphasis on writing improvement has a significant impact on student test scores in other disciplines, including science.

External Scoring

Another striking characteristic of the 90/90/90 Schools was frequent external scoring of assessments. While many schools continue to rely upon the idiosyncratic judgment of individual teachers for a definition of "proficiency," the high-achieving schools made it clear that no accident of geography or classroom assignment would determine expectations for students. Rather, these schools developed common assessment practices and reinforced those common practices through regular exchanges of student papers. One teacher would exchange papers with another teacher; principals would exchange papers with another school; and in one of the most powerful research findings, principals would take personal responsibility for evaluating student work.

When teachers exchange papers, it is imperative that they have a uniform basis on which to evaluate student work. The degree of agreement among teachers in their use of performance assessment scoring can be measured by "inter-rater reliability." Reliability, when the term is applied to traditional tests, is a measure of consistency. In the case of measuring consistency in scoring, it is simply the percentage of teachers who score an identical piece of student work the same way. If, for example, ten teachers evaluate a piece of student work, and eight believe that the work is "proficient" and two believe that it is only "progressing," then there is an 80 percent reliability rating for that test. This degree of reliability—80 percent—is the target at which teachers should aim as they jointly evaluate student work. It is very unusual (but not unheard of) for that level of agreement to be achieved the first time that teachers jointly score student work. More frequently, there are disagreements

among teachers on the evaluation of student work. These disagreements usually stem from one of two causes. First, teachers frequently use implicit scoring criteria that are not part of the official scoring guide. Examples of implicit criteria include such statements as "He should have written in cursive," or "She knew that she should have included that character in her essay." While these expectations may have been reasonable to these teachers, those criteria did not appear in the scoring guide. It is therefore little wonder that other teachers, who did not share those implicit expectations, failed to mark students down for these failings.

The second cause of teacher disagreement is the lack of clear specifications in the scoring guide itself. Too frequently a disagreement among evaluators leads to an argument rather than to an exploration of how agreement can be achieved through a revision of the scoring guide. "If we change the definition of proficient from this to that, perhaps we could agree on how to mark this paper." Words such as these are the basis of a far more meaningful discussion than, "Of course it's proficient! Don't you see?"

Long-Term Sustainable Results without Proprietary Programs

One of the most powerful findings of the 90/90/90 study is the continuous nature of the success of these schools, even as the poverty of students attending these schools remains intractable. Several of the schools listed below have consistently appeared on the 90/90/90 list, even as students change from year to year, as the effects of poverty grow more onerous, and as parents participating in welfare reform programs are less likely to be at home before and after school. Moreover, these schools are achieving their success without proprietary programs. Let there be no doubt: Our role in this research is as researcher and reporter. None of the 90/90/90 Schools used a specific "program" or any other proprietary model in order to achieve their success. On the contrary, we observed effective teachers and administrators using strikingly similar techniques without the assistance of externally imposed methods of instruction. The techniques used by these schools are replicable, but there is certainly not a need for schools to purchase special textbooks, curriculum materials, or secret information to achieve the level of success enjoyed by these schools.

Non-Proprietary Instructional Practices

In an era in which school leaders appear to engage in a perpetual quest for the magic bullet of educational success, it is noteworthy that none of the 90/90/90 Schools relied exclusively upon a proprietary program to achieve their success. Instead, these schools used consistent practices in instruction and assessment, with support

from local teachers. For those who believe that education remains an interactive process that cannot and should not be "teacher-proofed," these research findings are encouraging. The other edge of this particular razor is that we cannot depend upon proprietary systems to save us. It is the collective work of teachers, students, parents, and leaders that will ultimately lead us out of the present malaise. Every one of the 90/90/90 Schools had academic content standards, but so do many ineffective schools. The distinguishing characteristic of the 90/90/90 Schools was not merely that they had standards, but rather, how the standards were implemented, monitored, and assessed.

Data from the "90/90/90" Studies

A current list of some of the 90/90/90 Schools from Milwaukee, Wisconsin, is provided by the school system in their comprehensive accountability report. Since the publication of the first list in 1998, the number of schools qualifying for the designation has more than tripled. The data were independently verified by Schmoker (2001) in direct interviews with Milwaukee administrators. These schools have graciously hosted hundreds of visitors in the past few years as their successes have become more widely recognized. Researchers and educators should always be willing to share their sources of information and welcome the reviews of colleagues in the field. However, I cannot help but note how profoundly disturbing it is to me that I am frequently requested—demanded is not too strong a term—to produce the names and locations of these schools. In fact, these schools have received significant public attention through the Video Journal of Education, Volumes 802 and 803 (Linton Professional Development Corporation, 1998). Research should, of course, be subject to verification and scrutiny. Nevertheless, I cannot avoid noticing that in my many years of conducting, writing, and reviewing educational research, I have never seen such a demand for "names, dates, and places" accompany the allegation that children who are poor and children of ethnic minority groups perform badly on tests. When *The Bell Curve* (Herrnstein and Murray, 1994) was published with the widely accepted assertion that children who are black and poor perform badly on academic achievement tests, I cannot recall a single instance of demands for the names of students who were subjects of the studies cited. When we have demonstrated that poor and black children perform well, we are inundated with demands for verification. These demands speak volumes about the expectations of children based on their appearance and economic status.

After the original accountability report documenting the 90/90/90 Schools, Milwaukee Public Schools has issued subsequent accountability reports. The findings from these reports are striking. In brief, these findings include the following:

1. Techniques used by the 90/90/90 Schools are persistent. The students are still poor and their economic opportunities have not improved. Nevertheless, more

than 90 percent of the students in these schools continue to meet or exceed state standards.

2. Techniques used by the 90/90/90 Schools are replicable. The first time the district tracked these schools, only seven 90/90/90 Schools were identified. In the most recent report, 13 schools meet the criteria for this distinguished label.

3. Techniques used by the 90/90/90 Schools are consistent. These schools are not lurching from one fad to another. While they differ in some respects with regard to implementation, they are consistent with regard to the following areas of emphasis:

   • Writing—students write frequently in a variety of subjects.

   • Performance Assessment—the predominant method of assessment is performance assessment. This does not mean that these schools never use multiple-choice items. However, it is performance assessment in several different disciplines that local observers have associated with student progress.

   • Collaboration—teachers routinely collaborate, using real student work as the focus of their discussion.

   • Focus—teachers in these schools do not try to "do it all" but are highly focused on learning.

Additional Information on Success in Challenging School Environments

Over the years, I have continued to hear doubts and challenges that poor students can perform well. Indeed, the charge is frequently leveled that comprehensive accountability systems are disadvantageous for poor schools. In fact, systematic research from comprehensive accountability systems allows us to document and celebrate the success of students in these schools. Two additional sources of research on this subject come from strikingly different sources. Casey Carter, author of the "No Excuses" case studies from the Heritage Foundation (1999), provides a conservative viewpoint. The details of these cases are available at www.heritage.org. A politically liberal viewpoint is often associated with Kati Haycock and the Education Trust (1998, 2001). Their landmark research on student success in high poverty schools makes a striking case that these schools are not isolated anecdotes. Indeed, the fundamental finding from the Education Trust studies is that however important demographic variables may appear in their association with student achievement, teaching quality is the most dominant factor in determining student success. It turns out, of course, that teaching quality and subject matter certification are much more likely to occur in economically advantaged schools.

The case made by Haycock and others at the Education Trust is clear: the key variable is not poverty, but teaching quality. While poverty and other demographic variables may be important, they are not determinative in predicting student success. The detailed research from the Education Trust, including an interactive program allowing the user to specify the characteristics of a school and find specific data on comparable high-performing schools throughout the nation, is available at www.edtrust.org.

The consensus of the evidence from very different perspectives is clear: effective teaching and leadership make a difference. The lessons of the 90/90/90 Schools as well as the lessons of other studies provide convincing evidence that accountability systems, properly designed, can provide a wealth of information for those desiring to find the keys to improved achievement for all students.

Using the 90/90/90 Practices to Improve Achievement and Close the Equity Gap

Researchers and practitioners must always confront the gap between theory and reality, between anecdote and evidence. "Sure it worked there," the skeptics say, "but our kids are different." The ultimate test of the 90/90/90 research is whether it is sustainable and replicable. Simpson (2003) provides compelling evidence that the practices of the 90/90/90 Schools can be applied in a diverse urban environment with similar results:

> Like the city, Norfolk Public Schools, the first public school system in Virginia, has seen its fortunes go up and down. It's an urban district that serves a diverse population: 67 percent of students are black and 28 percent are white. More than 65 percent of students qualify for free and reduced-price lunches.
>
> - 100 percent of our schools met the state benchmarks in writing in all grades tested.
>
> - 100 percent of our high schools met the state benchmarks in chemistry.
>
> - 100 percent of our middle schools are fully accredited in earth science.
>
> - 100 percent of our middle and high schools showed positive trends in reading, literature, and research.
>
> Also, our schools reduced the achievement gap between white and black students in third, fifth, and eighth grades, with both groups continuing to improve. They decreased disciplinary actions by 15 percent, the number of long-term suspensions by 14 percent, and the number of expulsions by 66 percent. In addition, we have two "90/90/90 schools." These are schools with more than 90 percent of students eligible for free and reduced-price lunch, more than 90 percent are minority students, and more than 90 percent of

students met high academic standards on the state's Standards of Learning tests. (Simpson, 2003, pp. 43-44).

At the beginning of the 2002-2003 school year, I examined the accountability reports of each of the schools in Norfolk, Virginia, and conducted numerous site visits and interviews. In particular, I wondered if the buildings that experienced gains of 20 percent or more in their academic achievement in language arts, mathematics, science, and social studies were significantly different than their counterparts in other schools. The schools with the greatest gains were not similar demographically, as they included high-poverty and low-poverty student populations. The financial support, staffing patterns, union agreements, and central office support were similar for all schools. Therefore, neither the demographic variables of students nor the external variables of funding and labor agreements could explain the extraordinary differences between the schools. The keys to improved academic achievement are professional practices of teachers and leaders, not the economic, ethnic, or linguistic characteristics of the students. The Norfolk accountability system revealed striking similarities to other research on the characteristics of successful schools. Although surely there are many other traits shared by effective organizations of all types, the Norfolk Accountability System provided an insight into measurable indicators that were linked to the largest gains in student achievement. These characteristics also make clear that successful accountability is not the exclusive domain of the "Department of Accountability" in the central office, but rather is a responsibility shared throughout the system on many levels. The observations made on the basis of this inquiry are strikingly similar to observations I have made in other school systems over the course of several years. The following paragraphs highlight the nine characteristics that distinguished the schools with the greatest academic gains.

The Impact of Collaboration

First, the schools devoted time for teacher collaboration. This was not merely an exercise in idle discussion nor at attempt to get along in a friendly and collegial fashion. Rather, collaboration meetings were focused on an examination of student work and a collective determination of what the word "proficiency" really means. At first, teachers identified wide variations in their opinions and were alarmed to see how differently they evaluated the same piece of student work. In the course of many sessions, the most effective schools made time for collaboration very frequently and in some cases did this every day. Where does the time come from for effective collaboration? None of these schools had extra money in the budget or more hours in the day. Rather, they used the time that they already had with an intentional focus on collaborative scoring of student work. For example, the principals made their faculty meetings "announcement-free zones." Rather than drone through a laundry list of announcements (with inevitable comments and controversies), their rule was that the transmission of information would always be in writing. This allowed time formerly devoted to faculty meeting announcements

to be dedicated to collaboration. The principals were literally on the same side of the table as their faculty members, with faculty members who were experienced in collaborative scoring taking turns facilitating faculty meetings. The other source of time for collaboration was professional development meetings. Rather than presentations by outside staff developers, a significant degree of the professional development time was allocated to collaborative scoring. These educators knew that collaboration is hard work. Moreover, they understood that it is a skill acquired over time. Hence these remarkably effective schools did not have a "collaboration day" or a "collaboration workshop" but rather made the collaborative scoring of student work a part of their regular routine.

The Value of Feedback

Second, the schools with significant improvements provided significantly more frequent feedback to students than is typically the case with a report card. Emulating their most successful colleagues in music and physical education, teachers provided feedback in real time. They knew that a basketball coach does not provide hints on an effective jump shot nine weeks after an error, nor does a great music teacher note the improper position of the violinist's left hand weeks after noticing the mistake, but rather coaches and musicians provide precise and immediate feedback. In some cases, teachers took a triage approach, providing traditional report cards to successful and self-directed students, while providing weekly reports on their progress to students who were struggling. Their approach to feedback was consistent with Robert Marzano and his colleagues whose meta-analysis of research on student achievement revealed that feedback had a profound impact on student achievement, provided that the feedback was timely, accurate, and specific (Marzano, Pollock, and Pickering, 2001). The emphasis that these teachers placed on accuracy in feedback was remarkable. Unlike the "positive distortion" that clouds so much classroom feedback (Foersterling and Morgenstern, 2002), teachers with large gains were committed to feedback that was consistently accurate, with student performance compared to unambiguous expectations.

The Impact of Time

Third, the schools with large gains made dramatic changes in their schedule. Although they had the same budget, state requirements, teacher's union contract, and other restrictions as other schools in the system, the schools with large gains made remarkable schedule changes. At the elementary level, they routinely devoted three hours each day to literacy, with two hours of reading and one hour of writing. At the secondary level, they routinely provided double periods of English and mathematics. This was not a shell game in which they used the block schedule to double up some times but cut back on English and math in other times, but rather represented a genuine increase in instructional hours of math and English. The

essential nature of instructional time is hardly a new idea, yet in an astonishing number of schools, the schedule is revered more than the Pledge of Allegiance, Constitution, and Magna Carta combined. To break the mold in student achievement, these schools discovered, they had to break the schedule. It is interesting that this commitment to time for literacy instruction occurred in a state in which social studies and science content examinations were required. These teachers and principals did not change the schedule to over-emphasize literacy because they disregarded science and social studies, but rather because they knew that literacy was essential for success in every content area.

Action Research and Mid-Course Corrections

Fourth, teachers engaged in successful action research and mid-course corrections. In many of the schools with the greatest gains, their school accountability plans were not static documents set in concrete before the beginning of the school year, but dynamic and flexible guides. They asked the central office for permission to change goals and strategies that were not effective and start new ones that held promise, even during the school year. Moreover, these faculties and leaders learned from one another. An illustration of their commitment to the application of action research is the use of word walls at the secondary level. Because both the school improvement data and the instructional techniques associated with those improvements are transparent in a system of holistic accountability, the teachers who had achieved great things with students were subject to being questioned by colleagues throughout the system about their success. When in earlier years, elementary educators reported that significant improvements in vocabulary and reading comprehension results were associated with the implementation of word walls, the secondary science and social studies educators decided to adopt the idea. They created walls with words containing essential science and social studies vocabulary, sometimes associated with vivid visual images, and used those vocabulary words throughout the year. In other examples of effective action research, teachers replicated one another's writing rubrics, interdisciplinary assessments, and student motivation practices.

Aligning Teacher Assignments with Teacher Preparation

Fifth, principals made decisive moves in teacher assignments. Some writers have argued that when test scores are down, the entire school should be reconstituted and the entire faculty dismissed. In my observations, however, principals have made impressive gains by reassigning teachers to different grades within the same school. Consider what has happened to the curriculum—particularly in the fourth, fifth, and sixth grades—over the past decade. There has been an enormous growth in the

complexity of the curriculum, particularly in math and science, with an accompanying set of assumptions about the undergraduate curriculum of the teachers responsible for those grades. Those assumptions have sometimes been wildly inappropriate. When the fourth grade curriculum requires an understanding of algebra and scientific inquiry and the teacher's undergraduate preparation does not include those subjects, there is a challenge that will not be solved with a one-day staff development course in academic standards. The teachers whose undergraduate backgrounds fail to match the standards are not bad people nor are they unprofessional educators. Rather, their preparation is better suited to a different grade level. Effective leaders know that they should seek not to "fix" the person, but rather find a job (and accompanying set of standards) that best meets the teacher's abilities and backgrounds. By making decisive moves in teacher assignments, these principals saved the careers of some teachers and dramatically improved the achievement of their students.

Constructive Data Analysis

Sixth, successful schools included an intensive focus on student data from multiple sources, and specifically focused on cohort data. They were less interested in comparing last year's fourth grade class to this year's fourth grade class (which are, in most instances, different children) and more interested in comparing the same student to the same student. Their most important questions were not, "Is this year's class different from last year's class?" but rather:

- "What percentage of a group of students is proficient now compared to a year ago?"

- "What percentage of our students have gained one or more grade levels in reading when we compare their scores today to their scores a year ago?"

- "Of those students who were not proficient a year ago, what percentage are now proficient?"

- "Of those students who were proficient a year ago, what percentage are now advanced?"

In brief, these teachers compared the students to themselves rather than to other groups of students. This analysis allowed them to focus their teacher strategies on the needs of their students and not on generic improvement methods.

Common Assessments

Seventh, the schools with the greatest improvements in student achievement consistently used common assessment. This is a dangerous recommendation to

18

consider in an era in which the most frequently heard complaint across the educational landscape is that students are over-tested. To be sure, many students are over-tested; but they are under-assessed. The distinction between testing and assessment must be clear. Testing implies an end-of-year, summative, evaluative, process in which students submit to a test and the results—typically many months later—are used by newspapers and policy makers to render a judgment about education. By the time the results are published, they are ancient history in the eyes of the student and teacher. Contrast this to the best practice in assessment, in which students are required to complete a task and then very soon—within minutes, hours, or days—they receive feedback that is designed to improve their performance. Effective assessment is what great music educators and coaches routinely provide to their students. Moreover, great educators use assessment data to make real-time decisions and restructure their teaching accordingly. The track coach, for example, does not use the previous year's data to make decisions about assembling relay teams or selecting students to compete for the state finals. Rather, the most recent data available is far more important than the final results from the previous year. Similarly, the data from last quarter on a school-based assessment is far more helpful than the data from last year's test. Common assessments also provide a degree of consistency in teacher expectations that is essential if fairness is our fundamental value. Although individual teachers must have discretion on a day to day and hour to hour basis to teach, re-teach, and otherwise meet the needs of individual students, they do not have the discretion to presume that their students "just can't do it." The use of a common assessment for each major discipline allows for a combination of daily discretion and independence by teachers, while preserving a school-wide commitment to equity and consistency of expectations.

The Value of Every Adult in the System

Eighth, these remarkably successful schools employed the resources of every adult in the system. In holistic accountability systems, we can explore the extent to which professional development is distributed among all adults in the system. In a few remarkable cases, for example, there is profound respect for every employee, including bus drivers and cafeteria workers. The respect for these employees is evidenced by their inclusion in professional development opportunities in classroom management and student behavior. Leaders recognized that the student's day does not really begin in the classroom, but on the bus or perhaps during free breakfast. By committing their systems to consistency in the education and behavior of adults, these leaders ensure that every adult leader, from the bus driver to the food service employee to the classroom teacher is regarded as a significant adult leader in the eyes of students. The language concerning student behavior, sanctions, and rewards, is consistent and the results are impressive. Concomitant with gains in student achievement, these schools witnessed dramatic improvements in student behavior, including a reduction of bus misbehavior and disciplinary incidents outside the classroom.

Holistic accountability (Reeves, 2001) reviews allow a consideration of other extraordinary performances, including those by school nurses, library/media center specialists, school secretaries, custodians, counselors, psychologists, security guards, and many other unsung heroes whose exceptional efforts are disregarded in the typical accountability report. While holistic accountability does not provide a cookie-cutter approach to school success, it does reveal the remarkable impact of every adult in the system on student achievement.

Cross-Disciplinary Integration

Ninth, there is explicit involvement of the subjects that are frequently and systematically disregarded in traditional accountability systems—music, art, physical education, world languages, technology, career education, consumer and family education, and many other variations on the these themes. Analysis of holistic accountability data reveals that the involvement of these seemingly peripheral subjects in academic achievement is neither serendipitous nor insignificant. Rather, there is a deliberate strategy of involvement in these subjects in the improvement of academic results for all students. A few examples will serve to illustrate the point. Teachers meet to review student achievement data at a deep level, including the sub-scale scores. The discussion is not that "math scores are low" but rather that "the sub-scales reveal that we need to work in particular in fractions, ratio, and measurement." This leads the music teachers to develop activities in which musical rhythms reveal the relationship of whole-notes, half-notes, and quarter notes. Art teachers work on perspective and other representational art that makes explicit use of scale. Physical education teachers allow students to choose to run either a millimeter or a kilometer, and when they make the wrong choice, it is a lesson most students remember well.

In a striking example of collaboration in Norfolk, the teachers in music, art, and physical education collaborated to teach a social studies unit about African studies and the nation of Mali, the home of many of the students' ancestors. Using dance, literature, vocabulary, geography, history, song, and other engaging activities that crossed disciplinary boundaries, the teachers took the Mali unit out of the shadows of the final week of school and infused it throughout the school year. It is hardly an accident that these students also displayed astonishing improvements in their performance on state social studies tests.

Other Urban Success Stories

Norfolk is hardly an isolated example of success in urban school systems. In Indianapolis, Indiana, the Wayne Township Metropolitan School Corporation is among many that has demonstrated that academic improvement is compatible with high percentages of minority and poor students in the student body. In St. Louis,

Missouri, Dr. Chris Wright and her colleagues have led successful initiatives in both Riverview Gardens and Hazelwood school districts. Now, under the leadership of Dr. John Oldani and Dr. Dennis Dorsey of the Cooperating School Districts of St. Louis County, these techniques are having an impact throughout the St. Louis area.
— In Los Angeles County and Orange County, California, urban, suburban, and rural school systems are collaborating to create significant gains in student achievement.

The Wayne Township results are particularly interesting, as they represent not only an example of successful accountability, but also the ability of a complex urban school system to replicate the success of other systems. The Wayne Township experience demonstrates that holistic accountability is not merely the result of idiosyncratic case studies, but rather the result of systematic replication of best practices from within and outside a school system. The demographic characteristics of Wayne Township might be those of any urban system, with 26 different languages spoken by the students, free and reduced lunch enrollment as high as 80 percent in some schools, and minority enrollment increasing in a number of schools to the point that a majority of students are from minority ethnic backgrounds in some buildings. What is unusual, however, is the relentless focus of this school system on collaboration, academic standards, and nonfiction writing at every level. In particular, the years from 1999 through 2003 represent an extraordinary effort to augment the state's accountability system with a district-based holistic accountability system. In addition to the state tests, the district administers pre- and post- tests for every student in the fall and spring of each academic year. For the year ending in June of 2002, every single school made significant gains in mathematics and language arts. In addition, the schools with the highest poverty levels made the greatest gains, perhaps because those schools displayed the most intensive focus on changing schedules, instructional practices, building-level assessment, and leadership. It was therefore no surprise that when the state tests were administered in the fall of 2002, every building displayed significant growth, but those buildings with the highest poverty levels displayed the greatest growth in academic achievement. These gains exceeded 20 percent in the case of several schools within the district.

Without a constructive accountability system, these results might be passed off as the temporary reaction to test preparation resulting from pressure from state authorities. The facts contradict such a presumption. Every school in Wayne Township tracked specific practices in leadership and teaching. In the case of those schools with the greatest gains, there were common assessments on a monthly or quarterly basis. In addition, faculty meetings and staff development sessions were routinely devoted to collaborative scoring of student work. Each of the schools had
— common scoring rubrics so that there were consistent descriptions of what the word "proficient" means in practice. Following the lead of the district, each school embraced the use of "power standards" so that teachers were able to focus on a few of the most important standards rather than every single standard established by the state. This is among the most important observations of this holistic accountability study: higher test scores resulted not from mindless test prep and frantic coverage of

every standard, but rather from the thoughtful application of the most important standards to creative and engaging teaching strategies.

It was noteworthy that the schools that had the greatest gains did not eliminate special area courses, such as music, art, physical education, and technology. Rather, these courses were explicitly a part of the academic preparation of every student. In schools with the highest gains, each teacher in the special areas was given the standards in mathematics and language arts in which students needed the greatest amount of help. Each of these teachers incorporated some of those language arts and math standards into their daily lessons.

Finally, the principal was personally involved in the evaluation of student work. The building leader regularly met with students and parents to discuss student achievement in specific terms. Moreover, the principals personally administered common assessments every month in language arts and math. By giving up faculty meetings, the principal helped to provide additional time for collaborative scoring of student work. The principal also encouraged every teacher to display proficient and exemplary student work in a highly visible manner. The result of these displays was that every student, parent, and teacher had a clear and consistent understanding of what the school-wide scoring rubrics meant in practice.

The Impact of Holistic Accountability on Equity

As impressive as the improvements in academic achievement were in Wayne Township, the gains in equity were nothing short of extraordinary. Figure 19.1 showed the typical negative relationship between poverty and student achievement. The more likely a school is to have high percentages of poor and minority students, the less likely the school is to have a high proportion of the students achieve academic proficiency.

The line extending from the upper left to the lower right shows that as the percentage of students in poverty (as defined by those eligible for free or reduced lunch) increases, the achievement (as measured by test scores) decreases. This relationship is not perfectly negative (-1.0) but it is substantial in most national research, ranging from -.6 to -.9. The prevailing assertion in more than four decades of research on the topic is that variables such as student poverty account for 90 percent or more of the variation in student test scores (Marzano, 2003). If we stop with a consideration of Figure 19.1, then these prevailing assertions will carry the day. The accountability evidence, however, suggests that there are specific teaching, leadership, and curriculum strategies that will mitigate the impact of poverty.

Figures 19.2 through 19.5 indicate that the negative relationship between student poverty and student achievement is not a certainty. Although the grade 6 language arts scores are disappointingly negative (- .35), in both grades 3 and 6, the relationships between poverty and achievement are far lower than is the case

nationally, and in three out of four examples, the relationships are almost flat. In other words, this school system has demonstrated that the relationship between poverty and student achievement can be negligible.

Figure 19.2

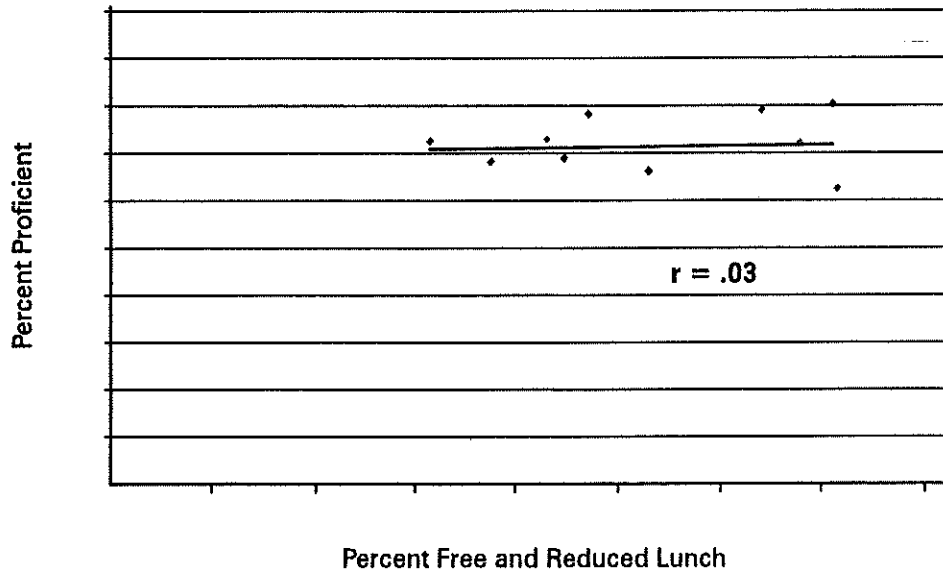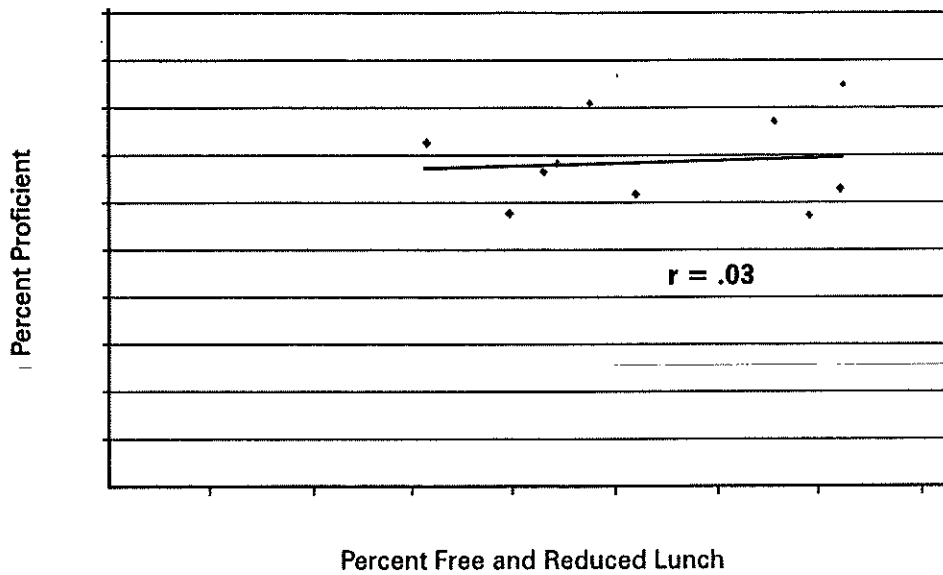## Relationship Between Poverty and 3rd Grade Language Arts Achievement

$r = .03$

Percent Proficient

Percent Free and Reduced Lunch

Figure 19.3

Relationship Between Poverty and 3rd Grade Mathematics Achievement

$r = .03$

Percent Proficient

Percent Free and Reduced Lunch

Equity Need Not Be A Dream

The Wayne Township experience demonstrates that equity need not be a dream. Every single building in the district—elementary through high school—achieved one of the following two equity indicators: The difference between students eligible for free and reduced lunch and the average was less than 10 percent, or the difference between the largest minority group of students and the average was less than 10 percent. These data points are totally consistent with the improvements in equity in Milwaukee, Freeport, Riverview Gardens (St. Louis metropolitan area), and others.

While no one disputes that poverty, linguistic differences, and culture can be important variables influencing student achievement, the research is clear that variables in teaching, curriculum, and leadership are profoundly important. In fact, these variables, that teachers and leaders can control, are more influential over student achievement than the intractable variables of poverty, culture, and language.
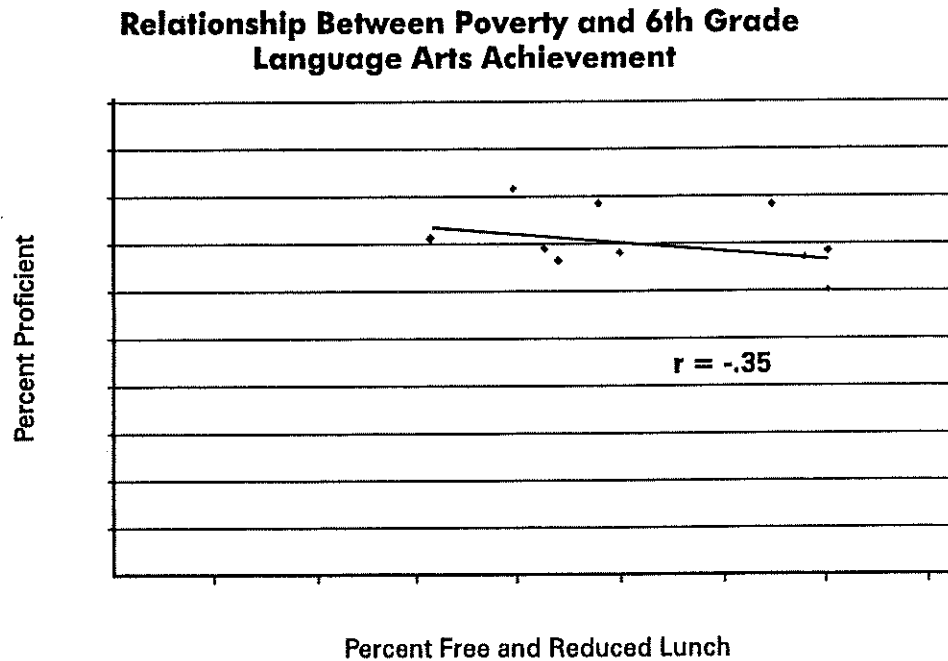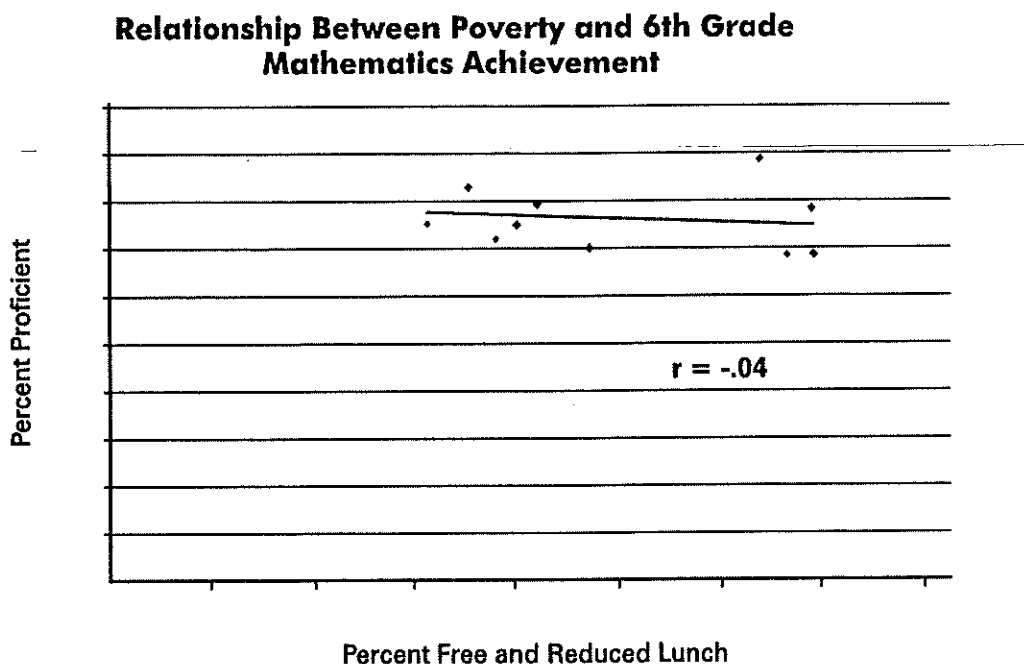
Figure 19.4

Relationship Between Poverty and 6th Grade Language Arts Achievement

$r = -.35$

Percent Proficient

Percent Free and Reduced Lunch

Figure 19.5

Relationship Between Poverty and 6th Grade
Mathematics Achievement

Percent Proficient (y-axis)

$r = -.04$

Percent Free and Reduced Lunch

Critics, Cynics, and Urban Education Success

We must take a few minutes to address the inevitable critics who appear to be constitutionally unable to believe that a success story in urban education exists. Whenever I share results such as those in Norfolk, Wayne Township, Milwaukee, Riverview Gardens, Freeport, or other successful urban schools, critics inevitably roll their eyes and allege that this surely must be a flash in the pan, the product of a frenzy of test preparation rather than sustainable reform. Others have claimed that the results must be due to the exclusion of under-performing children on test day. Still other critics claim that the students and teachers must be engaged in a massive cheating conspiracy. Others take issue with the methodology of the research, particularly if careful research controls (such as mobility and attendance) are used. The presence of those controls inflates achievement, the critics charge. After all, the studies reflect students who actually attend school. Of course, the absence of those controls would lead to charges of sloppy research. Either way, the critics find a way to ignore the continuing pile of research, of which my studies represent only a few pebbles. Marzano (2003) has assembled the most impressive evidence, using meta-analytic techniques that indicate the importance of teaching, curriculum, and leadership relative to poverty and ethnic identity. Demographic characteristics are relevant, but the preponderance of the evidence indicates that these characteristics are not destiny when it comes to academic achievement. The following is a brief consideration of challenges that I have heard made to the 90/90/90 research:

The only measure of success in this study is test scores, and there are better ways to assess student achievement. Test scores are a way, but by no means the only way, to assess student achievement. It is interesting that one of hallmarks of the 90/90/90 Schools was an unwillingness to tolerate annual state or district tests as the sole measurements of achievement. These schools consistently elevate the importance of classroom-based, teacher-made tests that are collaboratively scored and used to provide immediate feedback to both students and teachers. From a research and policy perspective, however, it is necessary to have some consistent data in order to understand student achievement. While accountability should indeed be a holistic endeavor with multiple assessments of achievement, common tests of literacy and mathematics are useful to evaluate student achievement over time. Finally, the best accountability systems, including the one used in the original 90/90/90 research, included a balance of state, district, and school-based measures. Moreover, it included a narrative report from each school, providing a balance of qualitative observation and quantitative data.

The excessive time devoted to reading means less time for science and social studies. This is true. Schools in the study were required by state law to take science and social studies tests, yet they made a deliberate trade-off to devote more time to reading comprehension and nonfiction writing, even if it meant that they had fewer hours of social studies instruction. This trade-off was wise for two reasons. First, their scores in social studies and science did not decline, but increased. One can speculate that it might have had something to do with the improved ability of students to read and understand the questions on the social studies and science tests. Second, our interviews of social studies and science teachers at the secondary level revealed their nearly unanimous conviction that the key to greater success in those disciplines at the secondary level was not more social studies and science instruction in elementary school, but students who could enter secondary school able to read on grade level. A substantial body of research (Foersterling and Morgenstern, 2002; Klentschy, Garrison, and Amaral, 2000) supports the teachers in this conviction.

The controls for attendance and mobility provide a positive bias for 90/90/90 Schools. This is not true. The accountability system provided "two-column" reporting for students in order to display the impact of mobility and attendance. In one column, the report shows the results for all students, and in the next column it shows the results for those students who were continuously enrolled during the school year. For attendance, the "all student" number was separated from the results for those students who attended school at least 90 percent of the time. These controls were made for all schools, not just the 90/90/90 Schools. Therefore, a parallel comparison was made to high poverty, high minority schools for students with good attendance and continuous enrollment, but who did not have the success of students in the 90/90/90 Schools. This is just good research design. In pharmaceutical research, we compare patients who receive the medicine (the experimental group) to those who receive a placebo (the control group). The research is only useful if those in the experimental group really take their medicine. If we are studying the impact of certain strategies in curriculum, teaching, and educational leadership, our research is of questionable value if we analyze the effects on students who were not present

for the curriculum, teaching, and leadership strategies. Finally, it was noteworthy that the schools that had high mobility (as defined by more than 80 percent of students taking the spring test not enrolled in September) and also high achievement, had strikingly similar characteristics to the 90/90/90 Schools, with an emphasis on writing and collaboration.

The 90/90/90 Schools used expensive programs, such as Success for All. This is not true. Some schools used Success for All, and others did not. This makes emphatically clear that the brand name alone of a literacy program is not the predictor of success, but rather the professional practices employed by teachers and leaders in the building. In fact, some Success for All schools had high results, while others had poor results. It was the replicable professional practices, not particular programs, that were associated with student success.

The effects are transient and dependent upon a particularly effective principal and faculty. This is not true. The effects are sustainable, with some schools maintaining this designation through different principals and high faculty turnover. The effects are replicable, with schools in other places (where there is also high turnover and teacher inexperience, particularly in high poverty schools). In the words of one teacher in the original study, "nobody volunteered to come to this school." Nevertheless, their collaboration, focus, and professional practices delivered results.

Conclusion

Perhaps the most compelling argument against any research about success in high poverty schools is the observation that there are cases where teachers are doing all of the right things, and yet student achievement remains low. There are no magic potions to deliver improved student achievement. The best that researchers and policymakers can do is to examine the preponderance of the evidence and draw appropriate conclusions. When a jury is presented with the evidence in a court case, it rarely has a perfect data set with unquestionable research. Rather, the jury confronts conflicting information, including information with errors, uncertainties, and differing interpretations. From this mix, we ask twelve people of good will and common sense to draw an appropriate conclusion based on the preponderance of the evidence.

The 90/90/90 research and the other evidence offered in this article fall far short of perfection. It does, however, contribute to the larger body of evidence that, in its totality, suggests useful strategies for high poverty schools. Moreover, in any research project, we must recognize that perfection is not an option. Rather, we can only choose among the errors that we commit, and attempt to minimize the risk of our errors. From a research perspective, we must choose between the risk of confirming a hypothesis that is not true and the risk of failing to confirm a hypothesis that is true. In the case of the professional practices recommended in this

article, we also have two potential errors. One error is the replication of these practices, including an increase in our commitment to literacy, nonfiction writing, and collaboration, and the subsequent discovery that the students really did not need all of that extra work after all. What is the risk of this strategy? Excessively literate students? Teachers who collaborate too much? The other error is the failure to act while we search for perfection or persist in a state of disbelief. Risks attendant with such delay will be debilitating for another generation of students. I do not claim that the 90/90/90 research and its many counterparts in the literature are perfect. I only suggest that the risks of this research being wrong are minimal. The risks if the research is correct and ignored are grave.

Questions for Discussion

1. What implications do the research findings on "90/90/90 Schools" have for your school or district? How would you implement these in your school or district?

COMMENTARY

Galileo's Dilemma

THE ILLUSION OF SCIENTIFIC CERTAINTY IN EDUCATIONAL RESEARCH

By Douglas B. Reeves

As he surveyed the heavens, Galileo made careful observations and challenged the prevailing hypothesis that the earth was the center of the universe. But this same scientist, so careful in his observations, also came to conclusions about the tides that were, by today's standards, laughably wrong.

That the scientific method can be both illuminating and wrong, even when practiced by a distinguished researcher, is a cautionary tale for educators, school leaders, and policymakers. As any observer of educational policy who has not been living in a cave knows, there are now federal mandates for the use of "scientific" programs in education. I know this because I have done what few members of Congress have done: I actually read the "No Child Left Behind" Act of 2001, signed into law by President Bush in January after passing Congress by overwhelming majorities. In its formidable 1,184 pages, the law uses the term "scientific" or "scientifically" 116 times and the word "research" 246 times. The current controversy over what "scientific research" means in the context of education implies a dichotomy between certainty and sophistry that exists only in the minds of partisans who appear to revel in yet another fact-free debate on educational policy. Let us separate myth from reality:

Myth No. 1: Science grants certainty. What does scientific research really mean? Does it, as its proponents imply, provide a world in which, if we only followed the salutary models of medicine, chemistry, and physics, rational people would agree on clear and obvious solutions? Or does it give us complexity and uncertainty, with debates over the effectiveness of mammography, the sequence of elements, and the number of planets in our solar system, along with the quandaries over tides and planetary bodies confronted by Galileo? When scientific methods are applied, researchers can disagree. They can even be wrong. Even in the hard sciences, controversies abound and certainty is elusive. While educators can learn much from scientific methods, the insinuation that these methods grant certitude is, to put it charitably, a hypothesis unsupported by the evidence.

Myth No. 2: Double-blind studies, such as those used in pharmaceutical research, are the gold standard for educational research. In pharmaceutical studies, the control group receives a placebo while the experimental group receives the real drug. In an astonishing number of cases, both groups show evidence of improved health. That is, something that researchers know to be valueless demonstrates an apparent impact on patient health. As a result, researchers do not have a clean line of demarcation between success and failure, but rather some evidence that some degree of health is associated with some dosage of the experimental medicine that is less evident in the absence of that drug. Where there is a relationship between the experimental medicine and improved health, researchers note that there is an association—a statistical correlation—between the drug and the condition of the patient. They cannot make conclusions about cause and effect until they have a detailed understanding of the physiology of the biochemical reactions caused by the medicine. Sometimes, as is the case with the origin of many cancers, correlation is all that scientists have, as the physiological evidence remains unavailable.

The most serious problem with pharmaceutical studies is that other variables, including the condition of the patient, nutrition, attitude, exercise, diet, sleep, and a host of other personal and environmental conditions that

> When scientific methods are applied, researchers can disagree. They can even be wrong.

Galileo's Dilemma

affect the medicine are not always perfectly controlled. Educational researchers, the presumed unscientific slugs in this debate, have not yet figured out how to control the nutrition, attitude, exercise, diet, and sleep of their research subjects, any more than they have figured out how to control the 18 hours each day spent outside of school.

Myth No. 3: The No Child Left Behind Act clearly defines scientific research. In fact, a reading of the plain language of the bill makes two inferences abundantly clear. First, the demand for scientifically supported programs, however pervasive, does not exist in a vacuum. Throughout the bill, the same sentence links a demand for such programs with an equally strong imperative for support of a broad and academically rigorous curriculum. In numerous instances, the same sentence links scientific programs with a demand for professional-development programs. Most state standards require that 4th grade students comprehend the logic of Venn diagrams, in which students must understand that a statement can be part of one set but not necessarily represent a definition of the entire set. Participants in the debate over educational research would be well advised to rise to this standard. To put a fine point on it, the assertion that "science equals phonics" is only true with respect to the fact that some research studies support the use of phonics as part of an effective reading program. The assertion that "any program that does not include phonics is not scientific" does not meet the standard of logic we require of 4th graders. Personal opinions, distorted case studies, and flimsy observations can all masquerade as "research."

Myth No. 4: Anything bearing the label "research" is worthy of the name. One need only recall the tobacco advertisements of the 1940s in which physicians endorsed the soothing effect of cigarettes on the throat to question the relationship between authority figures and putative research conclusions. Personal opinions, distorted case studies, and flimsy observations all masquerade as "research." Galileo's successors in the 19th and early 20th centuries used their version of science to prove the superiority of the Scandinavian over the Italian and the rightful subordination of the African to the European. Academic journals in the early days of the 21st century

> We do need scientifically based programs in education. But real science involves ambiguity, experimentation and error.

allow the inconveniences of sample size and detailed disclosure of experimental methods to give way to political agendas. Rather than be defensive, educators should acknowledge these problems, just as researchers in medicine, physics, and chemistry regularly air their dirty linen and, with equal amounts of clumsiness and rigor, advance the cause of reason.

In education, the mantras of "studies show" and "research proves" are the staple of too many vacuous keynote speakers for whom a footnote is a distant memory of a high school term paper. The real researchers I know confess that our work is but a pebble on a mountain of research begun by others, list the details of findings, welcome double-checks and criticism, and eat crow on a regular basis, firm in the conviction that transparent error is the price we pay for knowledge. Our mistakes involve more work and more risk than speculation unencumbered by evidence, and by our mistakes we simultaneously confess error and advance knowledge.

The best we can do is consider a variety of conflicting studies and recognize the inherent uncertainties of research. The frailties of scientific research do not render us helpless. We can formulate sound opinions and make well-reasoned decisions on the allocation of scarce resources based on the information available. Rather than asserting that we have found ultimate truth with as much conviction as Galileo had in his false conclusions about ocean tides, we can acknowledge our limitations. The best we can do is consider a variety of conflicting studies and recognize the inherent uncertainties of research. At the same time, we must challenge the "scientifically based" assertions of others, particularly when prejudgment is substituted for fact.

Congress and the president got it right, though perhaps not in the way that they intended. We do need scientifically based programs in education. But real science involves ambiguity, experimentation, and error. However distasteful that trio may be, it is far superior to political agendas, uninformed prejudice, and breathless enthusiasm for the flavor of the month.

Douglas B. Reeves is the chairman of the Center for Performance Assessment and the author of Holistic Accountability *(Corwin Press, 2002) and* The 20-Minute Learning Connection *(Simon & Schuster, 2001). He lives near Boston.*

HARVARD EDUCATION LETTER

Search HEL's site

go

Past Issues

March/April 2002

Accountability-based reforms should lead to better teaching and learning-period

By Douglas B. Reeves

"We have to think about accountability in a very different way," says Douglas B. Reeves, chairman and founder of the Center for Performance Assessment and the International Center for Educational Accountability. "We have done a splendid job of holding nine-year-olds accountable. Let me suggest as a moral principle that we dare not hold kids any more accountable than we expect to hold ourselves."

At a recent forum hosted by the Principal's Center at the Harvard Graduate School of Education, Reeves outlined the principles of what he believes comprise effective school-based accountability systems. His remarks were edited for this issue

Principle #1: Congruence

Objectives and strategies are sometimes developed in complete contravention to what the accountability system calls for. Accountability must be the unifying theme that draws strategy, rewards, recognition, and personnel evaluations together.

I once worked in a district that planned to focus its accountability system on the principle of prioritized standards--that is, focusing on the most important standards rather than trying to cover everything at once. That was the rhetoric. But the first line of the teacher evaluation form was, "Did the teacher cover the curriculum?"

In another district, accountability-minded educators said, "We look at the evidence. We know that if more children are involved in extracurricular activities, our attendance and student achievement will be better." Yet when you looked at the recognition and reward system in that district, the teachers they rewarded and recognized were the most exclusive, the ones who protected their classes and their extracurricular activities from any students other than the cream of the crop. In both of these cases, the accountability system was contradicted by the objectives and strategies.

Principle #2: Specificity

If I go to one more conference where we hold hands and chant, "All children can learn," I'm not going to be able to take it anymore. I believe that all children can learn, but I have never achieved anything with a mantra. Accountability is not about chanting mantras; it's not about generalities.

31

We've got to know specifically what works. We've got to investigate which strategies in our own communities are specifically associated with improved student achievement. And let's focus on behaviors, not just test scores--in other words, measure what the grownups do. We need to set as many standards for the adults--the board members, the administrators, the teachers, perhaps someday even the parents--as we do for kids.

Principle #3: Relevance

There ought to be a direct relationship between the strategies schools employ and improvements in student learning. Of course, relevance isn't always obvious. Some research indicates that, with the exception of attendance, the number-one factor associated with improved test scores and behaviors in the classroom happens to be more nonfiction writing.

It may be obvious that more nonfiction writing is related to better writing scores, and it may make sense that more nonfiction writing is highly related to better reading scores. Less obvious is the fact that even a little more nonfiction writing in a curriculum is also related to better math, science, and social studies scores. In these instances, we find specific relationships between our classroom strategies and our results.

Do these relationships prove causality? Not necessarily. But they do provide us with a way of testing the hypothesis that more nonfiction writing will improve test scores and student behavior. If I were to ask every teacher, "Why can't you do more nonfiction writing?" many would say, "I don't have the time." Time, time, time is the number-one issue. These teachers are articulating the hypothesis that if they spent more time on writing, they wouldn't be able to cover the curriculum, and that would make scores go down. I may not have been able to prove causality, but I have disproved that hypothesis.

Principle #4: Respect for Diversity

"All children can learn" does not mean "all children are the same." Furthermore, diversity is not merely about external characteristics. If we're really going to take this seriously, that means we start looking at diversity on the inside as well as diversity on the outside. Making this principle both a moral and an intellectual part of the curriculum will require taking different approaches in different schools. That is, it will require a diversity of approaches, diversity of techniques, and diversity of teaching strategies.

When U.S. Secretary of Education Rod Paige was running Houston's schools, he did not say, "My way or the highway" to 200 schools. He said, "You want respect for diversity, including different styles, approaches, and strategies? You got it. But the price of freedom is transparency. The price of trying different things is being able to come to one another, and come to me, with transparent results. Tell me what worked, tell me what didn't work." That's what accountability requires. You can embrace different strategies provided that you report those strategies. Win or lose, succeed or fail, we report them.

It's important to remember that respecting diversity doesn't mean anarchy or that all views are equal. You can have respect for diversity without giving up foundational principles. We have the ability, maybe even the mandate, to say that some values are better. The values of freedom, truth, and justice are better than the values of oppression and totalitarianism. That's the kind of thing we ought to be able to say. Not every principle is up for grabs.

Principle #5: Continuous Improvement

Jeff Howard, president of the Efficacy Institute, uses an analogy that may

resonate for people who have kids at home. He calls it the Nintendo Effect, which refers to the child who cannot focus or concentrate and is always moving about the classroom until you turn on the Nintendo machine, whereupon the child is transfixed, moving not a follicle of hair as he sits for hours in front of the machine.

The question Dr. Howard asks is, "How long would that child be staring at the screen if you mailed his Nintendo scores to him nine weeks hence?" Part of what keeps him engaged is not just what's happening on the screen. It is that he gets feedback that is timely, immediate, and relevant. If we're going to build a holistic accountability system, once-a-year feedback is not sufficient. We should be building a system that every month gives feedback to our children, our leaders, and our teachers so we can get busy building better instructional systems.

Principle #6: Focus on Achievement, not Norms

There is actually a state where the Board of Education voted that 80 percent of students must be above average. Now, I have taught statistics for a long time, and no amount of listening to Garrison Keillor will convince me that that is a possible distribution. But there's another issue here. When you hear these comparisons made to norms and you hear comparisons made to the average, normally the visceral reaction is that this is something that hurts poor kids. However true that might be, it also hurts advantaged kids.

The bell curve is insidious for all kids. It is an ineffective, inappropriate way to measure student achievement. You've got some "above average" kids who are inappropriately complacent and who are hurt by norms as surely as kids who are in the low end of the bell curve. Do you know the 55th-percentile kid who gets a 55th-percentile score in reading and cannot write an essay to save his soul? The 55th-percentile student in math who cannot apply the algorithm in different contexts? The only thing that really matters is whether students are meeting expectations that are clear, objective, and immutable--not who beat whom.

Douglas B. Reeves is chairman and founder of the Center for Performance Assessment and the International Center for Educational Accountability.

For Further Information

D.B. Reeves. *Accountability In Action: A Blueprint for Learning Organizations.* Denver: Advanced Learning Press, 2000.

D.B. Reeves. *Making Standards Work: How to Implement Standards-Based Assessments in the Classroom, School, and District.* Denver: Advanced Learning Press, 1997.

Center for Performance Assessment, 1660 South Albion St., Suite 1110, Denver, CO 80222; 800-844-6599; 303-504-9312 (local); fax: 303-504-9417; email: info@makingstandardswork.com.

EDUCATIONAL LEADERSHIP

Evaluating Administrators

Results of the National Leadership Evaluation Study reveal the urgent need for fair, specific, and constructive leadership evaluation systems.

Douglas B. Reeves

April 2004

I have bad news, worse news, and a bit of good news. The bad news: Education leadership evaluation is a mess. Our national survey of leadership evaluation instruments reveals an astonishing disregard for what we know about effective feedback and meaningful evaluation. Newspapers provide box scores purporting to evaluate superintendents and principals on the basis of student test scores, but we rarely see any analysis of the impact of leadership on teaching and curriculum (Reeves, 2002a, 2002b).

The worse news: Improving leadership evaluation will be difficult. As the continued use of the grading system and the seven-period high school schedule attests, schools tend to cling to long-standing practices despite mountains of evidence pointing to the need for change.

The potentially good news: A better way exists. School systems can reject their dependence on ambiguous, demoralizing, and destructive leadership evaluation systems. One alternative—Multidimensional Leadership Assessment—has the potential to transform leadership evaluation from a blight on the education landscape to a constructive instrument of education policy.

Ineffective Evaluation

Author Sebastian Junger defined *the perfect storm* in his book of the same name (1998) as one in which many different variables came together at the same time to create particularly destructive consequences. The Center for Performance Assessment, in our National Leadership Evaluation Study, found a "perfect storm" of failure: The acute and growing shortage of education leaders is accompanied by a leadership evaluation system that simultaneously discourages effective leaders, fails to sanction ineffective leaders, and rarely even considers the goal of improved leadership performance. We reviewed hundreds of leadership evaluation systems and studied thousands of pages of documents in search of an example worthy of emulation.

These leadership evaluation systems do not come from the pens of incompetent bureaucrats; they come from intelligent and thoughtful people. But in almost every case, the evaluation systems are deeply flawed. These systems tolerate mediocrity, fail to recognize excellence, turn a blind eye to abuses, accept incompetence, and systematically demoralize courageous and committed leaders. Despite the exemplary work of such groups as the Council of Chief State School Officers and its Interstate School Leaders Licensure Consortium (ISLLC) (1996), the reality of leadership evaluation remains far removed from the ideal.

Thanks to such educators as Danielson (2002) and Darling-Hammond and Sykes (1999), schools have made significant strides in transforming teacher evaluation standards into practice. Unfortunately, although more than two dozen states and many school systems claim to have adopted the ISLLC standards, many of their administrator evaluation systems fail to implement

those standards with any degree of precision. Although we are certain that examples of constructive and specific administrator evaluation systems exist, our failure to find them was not for lack of effort. The National Leadership Evaluation Study findings suggest that such stellar leadership evaluation systems are the exception rather than the rule.

Findings of the Study

We conducted the National Leadership Evaluation Study from March 2002 to September 2002 with a nonrandom sample of 510 leaders—including district superintendents, central office administrators, and principals—from 21 U.S. states. The average respondent had 11.4 years of total leadership experience and 4.9 years in his or her current position. In addition to conducting the survey, we studied the leadership evaluation instruments used by more than 700 schools. Here is a sampling of our findings:

- More than 18 percent of the leaders we studied had never received an evaluation in their current position. In the words of one of our research subjects, "The worst evaluation experience was no evaluation at all. The message was that I was not important enough for my supervisor to take time to give me an evaluation."

- Of the leaders who were evaluated, 82 percent found leadership evaluation to be inconsistent, ambiguous, and counterproductive.

- Fewer than half of the respondents (47 percent) agreed that their most recent leadership evaluation was related to student achievement.

- Only 54 percent of the leaders said that their evaluation was based on clear standards.

- Only 47 percent of the leaders said that their evaluation was sufficiently specific to help them improve their performance.

- The higher the level of leadership responsibility, the lower the satisfaction with leadership evaluation instruments. New administrators more frequently received helpful and constructive coaching and feedback. Evaluation was least helpful for veteran administrators and central office directors. Leadership evaluation was at its worst when school boards were evaluating superintendents.

The narrative comments from respondents in our study reveal feelings of anger, betrayal, and despair. For a nation that will lose about half of its current school leaders to retirement within the next eight years (Dipaola & Tschannen-Moran, 2003), the systematic demoralization of the current leadership pool is destructive and foolish. With the growing number of unfilled leadership positions and an alarming number of leaders leaving the field of education, it is time for fundamental reform. The United States needs a new form of leadership evaluation, and it needs it now.

Ambiguous Leadership Standards

The problem starts with the definition of *leadership*, particularly in the context of education. Our survey reveals that the expectations articulated in most evaluation systems are at best ambiguous. At worst, they are contradictory, impossible, and inconsistent with common wvalues and mountains of research.

Most of the evaluation systems we reviewed eschewed descriptive rigor in favor of education jargon. The following statements come from local, state, and national performance expectations for school leaders. Each statement is followed by a challenge that any leader being evaluated by such a standard would want to consider.

Expectation: "The administrator facilitates processes and engages in activities ensuring that curricular, cocurricular, and extracurricular programs are designed, implemented, evaluated, and refined."

Challenge: What in the world does this mean? How would we know if this standard has been met? Do evaluation and refinement refer to what is popular or what is effective?

Expectation: "Stays current with research and theory regarding motivation. Keeps abreast of the latest developments in the field of education."

Challenge: Any research and theory? Much of it could well contradict the goals and values of the school system. This goal appears to endorse a collection of fads; school leaders could fail to distinguish what is "current" from what is important, valid, tested, and trustworthy.

Expectation: "Provides information on curriculum/instruction."

Challenge: Is there a single school administrator who can fog a mirror who does *not* do this? The issue is not whether or not the leader provides information; rather, the issue centers on the quality of that information, and whether it will lead to good decisions and improve student achievement.

Expectation: "Facilitates processes and engages in activities ensuring that relevant demographic data pertaining to students and their families are used in developing the school mission and goals. Diversity is considered in developing learning experiences."

Challenge: Does this mean that good leaders have different goals for low-income schools than for affluent schools? Is it a good idea to develop different goals for schools on the basis of their ethnic composition? If the families have a culture of low expectations, should schools mirror those expectations?

Expectation: "Participates in professional development activities."

Challenge: My 4th grader's hamster can participate in professional development activities. What does this tell us about the impact of using new knowledge and skills to become a more effective leader?

Sometimes the expectations are internally contradictory. One district's leadership evaluation instrument requires its effective leader to "carefully weigh consequences of contemplated action"; a few sentences later, the evaluation assesses the same leader on whether he or she "is action-oriented; presses for immediate results" and "is decisive; doesn't procrastinate on decisions." The same evaluation form requires the leader to simultaneously "hold to personal opinions," "exhibit a need to control most situations," and "demonstrate adaptability and flexibility." We might gently suggest that an administrative certificate and a doctorate are not the criteria sought by this district, but rather some combination of divinity and multiple personality disorder.

Incoherent Leadership Evaluations

Not every leadership evaluation instrument that we examined is so deeply flawed in establishing clear leadership standards. But even those with clear standards often suffer from ambiguous descriptions of performance levels. Typical performance levels include "exceeds expectations," "meets expected performance levels," "superior," or "average"—without any clear indication of which specific leadership behaviors deserve such labels.

Without specification, the leader's rating on these performance levels depends on the idiosyncratic judgment of the evaluator. However wise and insightful an individual evaluator may be, these judgments are doomed to be inconsistent and practically useless for coaching. The person evaluated only knows that one evaluator regarded him or her as "outstanding," another evaluator believed that the same leadership traits and behaviors merited a rating of "meets standards," and yet a third evaluator said that the same performance "exceeds expectations." We should not expect leadership wisdom to emerge from such an ambiguous pool of linguistic slop.

Even if the standards themselves are clear, descriptions of performance devolve into the

36

linguistic quicksand of "sometimes" compared with "seldom," or "frequently" compared with "often," or "exceeds expectations" compared with "satisfactory." Intelligent people of goodwill can disagree about what any of these descriptions mean.

Perhaps the least-helpful performance ratings are such descriptions as "growth needed." This rating is invariably a negative comment in the context of evaluation, yet I strain to think of a single leader—from Alexander the Great, to Napoleon, to Churchill, to Eleanor Roosevelt, to Martin Luther King, Jr., to the best school leaders I have observed in more than a million miles of travel—who would not enthusiastically check the box next to "growth needed" when describing himself or herself. To put it bluntly, when is growth *not* needed? Presumably, when one is dead.

Effective evaluation systems enable both the evaluator and the one being evaluated to understand clearly the differences between various levels of performance. Michael Jordan, for example, was acutely aware of the difference between putting the ball in the basket and hitting the rim. His fans shared his perceptions of clarity in evaluation. Sarah Chang, along with the vast majority of her audience, knows the difference between an F-natural and an F-sharp. But do school leaders, to whom we entrust our children and billions of dollars in resources, know the difference between performance that is exemplary, proficient, and below expectations?

Authority/Responsibility Disequilibrium

We wish our leaders to be some mythical combination of folk heroes, in which they have the insight of Lao-tzu, the courage of a New York firefighter, and the work ethic of Paul Bunyan. In the real world of school leadership, however, the relationship between demands and authority leads to results that are more prosaic. This does not stop the developers of leadership evaluations from making the grand resumption that the school principal or district superintendent enjoys enormous powers.

The most glaring examples of the authority/responsibility disequilibrium occur when we hold education leaders responsible for the actions of others—ranging from the most recalcitrant employee to the most apathetic community member—even though they lack the authority to control the actions of either of these stakeholders. One set of leadership standards reviewed in our study, for example, required the leader to "ensure that staff and community understand the analysis of student data." Leaders can provide information to the staff and community and can even assess the staff's knowledge, but they cannot "ensure" understanding. Another leadership evaluation standard required the leader to "ensure a balanced budget." Meeting this standard might require controlling the local property tax rate, the price of oil in Iraq, the impact of hail on the roofs of schools in Kansas, or the number of snow days in Idaho. To put it gently, snow happens, along with a host of other natural events that affect the budget and are far beyond the control of school leaders.

The desirable outcomes in a school or district fall along a continuum: Some areas are subject to the leader's control; others are subject only to the leader's influence; and still others are beyond the leader's influence. For example, the leader may directly control the timing and content of a faculty meeting. He or she may directly influence the quality and content of teacher evaluations, working within the constraints of the collective bargaining agreement. He or she can only indirectly influence the quality of teaching in the classroom and the motivation of students. And the extent to which students received adequate prenatal care and early childhood education is far beyond the control or influence of most school leaders. An honest leadership evaluation system will specify expectations for leaders that are appropriate, recognizing the amount of influence or control that the leaders can exert over each area.

A Better Way

Schools need an alternative to the vacuous exercises now called leadership evaluation. A better model would provide specific, accurate, and timely feedback. Rather than an event that occurs

once a year (or in the case of senior leaders, every three or four years, always too late to influence performance), evaluation should consist of frequent feedback and provide multiple opportunities for continuous improvement. Rather than providing meaningless performance levels—such as "meets expectations," "above average," or "progressing toward standards"—the ideal leadership evaluation system would describe in specific terms the difference between distinguished performance and performance that is proficient, progressing, or failing to meet standards.

We have developed a model for more-effective leadership evaluation that meets these requirements and reflects best practices in performance assessment: the Multidimensional Leadership Assessment (MLA). The MLA model encompasses 10 dimensions of leadership, including resilience, personal behavior, student achievement, decision making, communication, faculty development, leadership development, time/task/project management, technology, and learning. This list, although hardly exhaustive, represents a compromise between the very extensive list of leadership requirements in such documents as the ISLLC standards and the vague assessments used in many school districts. For each dimension of leadership, we developed subcategories of specific leadership behaviors. Figure 1 (p. 54) shows the 10 dimensions and their subcategories.

Figure 1. Major Dimensions for Constructive Leadership Evaluation

1.R esilience
- o 1.1. Constructive reaction to disappointment and failure
- o 1.2. Willingness to admit error and learn from it
- o 1.3. Constructive management of disagreement with leadership and policy decisions
- o 1.4. Constructive management of dissent from subordinates
- o 1.5. Explicit improvement of specific performance areas after considering the previous leadership evaluation

2.P ersonal Behavior
- o 2.1. Integrity
- o 2.2. Emotional self-control
- o 2.3. Compliance with legal and ethical requirements in relationships with employees
- o 2.4. Compliance with legal and ethical requirements in relationships with students
- o 2.5. Tolerance of different points of view within the boundaries of the values and mission of the organization
- o 2.6. Organization, including calendar, desk, office, and building(s)

3.Stud ent Achievement
- o 3.1. Student achievement results
- o 3.2. Student achievement reporting to students, parents, teachers, and school leaders
- o 3.3. Use of student achievement data to make instructional leadership decisions

38

- o 3.4. Understanding of student requirements and academic standards

- o 3.5. Understanding of present levels of student performance based on consistent assessments reflecting local and state academic standards

- o 3.6. Decisions in teacher assignment, course content, schedule, and student curriculum based on specific needs for improved student achievement

4.Deci sion Making
- o 4.1. Factual basis for decisions, including specific reference to internal and external data on student achievement and objective data on curriculum, teaching practices, and leadership practices

- o 4.2. Clear identification of decision-making structure, including which decisions are made by consensus and which are made by the leader with advice from others

- o 4.3. Decisions linked to vision, mission, and strategic priorities

- o 4.4. Decisions evaluated for effectiveness and revised when necessary

5.Com munication
- o 5.1. Two-way communication with students

- o 5.2. Two-way communication with faculty and staff

- o 5.3. Two-way communication with parents and community

6.F aculty Development
- o 6.1. Understanding of faculty proficiencies and needs for further development

- o 6.2. Individual consideration of faculty needs linked to vision, mission, and strategic priorities

- o 6.3. Personal participation in leading professional development initiatives

- o 6.4. Congruence of strategic objectives and professional development content

- o 6.5. Recognition and rewards strategically linked to most-important faculty and staff behaviors

- o 6.6. Inclusion of faculty in decision making, including collaboration and advice on major leadership decisions

- o 6.7. Formal and informal feedback to colleagues with the exclusive purpose of improving individual and organizational performance

7.L eadership Development
- o 7.1. Strong assistant administrators who are capable of immediately assuming leadership responsibility in this school or other buildings

- o 7.2. Consistent identification of potential future leaders
- o 7.3. Evidence of delegation and trust in subordinate leaders

8.Ti me/Task/Project Management
- o 8.1. Consistent maintenance of daily prioritized task list
- o 8.2. Choices for time management focused on the most-important priorities
- o 8.3. Clear objectives and coherent plans for complex projects
- o 8.4. History of completion of projects on schedule and within budget

9.Te chnology
- o 9.1. Demonstrated use of technology to improve teaching and learning
- o 9.2. Personal proficiency in electronic communication
- o 9.3. Coherent management of technology resources, technology staff, and information

10. Learning
- o 10.1. Personal understanding of research trends in education and leadership
- o 10.2. Evidence of personal growth and learning

Source: Multidimensional Leadership Assessment, Center for Performance Assessment.

To make the dimensions and subcategories meaningful and useful for their specific needs, districts must develop detailed descriptions of leadership performance that range from exemplary, to proficient, to progressing, to not meeting standards. Figure 2 (pp. 56–57) provides an example of such a continuum of performance for three selected subcategories under the dimension of resilience.

Figure 2. Continuum of Performance for Selected Subcategories of Leadership Behavior

Dimension 1: Resilience				
Leadership Subcategory	Exemplary (Systemwide Impact)	Proficient (Local Impact)	Progressing (Leadership Potential)	Not Meeting Standards
1.1. Constructive reaction to	Public reports, including accountability	Readily acknowledges personal and	Acknowledges personal and organizational	Defensive and resistant to the acknowledgment

disappointment and failure	documents, plans, and oral presentations, include frank acknowledgment of prior personal and organizational failures and clear suggestions for systemwide learning resulting from those lessons.	organizational failures.	failures when confronted with evidence.	of error.
1.3. Constructive management of disagreement with leadership and policy decisions	Articulates disagreements with policy and leadership decisions and advocates for a point of view based on the best interests of the organization. Challenges executive authority and policy leaders appropriately with evidence and constructive criticism, but once the decision is made, fully supports and enthusiastically implements organizational decisions.	Accepts and implements leadership and policy decisions.	Sometimes challenges executive and policy leadership without bringing those concerns to appropriate executive and policy authorities. Sometimes implements unpopular policies unenthusiastically or because "I'm just following orders, but I don't like it."	Ignores or subverts executive and policy decisions that are unpopular or distasteful.
1.4. Constructive management of dissent from subordinates	Creates constructive contention, assigning roles if necessary to deliberately generate	Uses dissent to inform final decisions, improve the quality of decision	Tolerates dissent, but there is little dissent in public because subordinates do not understand the leader's	Dissent is absent as a result of a climate of fear and intimidation.

	multiple perspectives and consider different sides of important issues. Recognizes and rewards thoughtful dissent, thus conveying broader support for the final decision. Uses dissenting voices to learn, grow, and, when appropriate, acknowledge the leader's error.	making, and broaden support for final decisions.	philosophy about the usefulness of dissent.		

A specific, accurate evaluation system, such as the Multidimensional Leadership Assessment, enables leaders to understand clearly their expected performance and to engage in frequent reflection and self-correction. This system also enables school boards and senior leaders to make their expectations clear before a leader is hired. It encourages proactive evaluation—starting the evaluation process before the first day on the job rather than as a reaction to disappointing performance. Most important, this model offers clarity and consistency. Because the standards are explicit and the performance continuum is unambiguous, leaders receive consistent, fair, and constructive feedback.

Time for a Change

The past three decades have witnessed tremendous strides in every area of education assessment. It is time for us to apply those lessons to the assessment of education leaders.

If schools persist in using the current unfair, ambiguous, and demoralizing system of leadership evaluation, then a generation of education leaders will remain subject to the whims of personal preference and newspaper headlines. The accompanying burnout of leaders and national shortage of people willing to occupy key leadership positions will have grave consequences throughout the system.

The Multidimensional Leadership Assessment offers a fairer and more constructive alternative than do the vast majority of existing leadership evaluation systems. While we await studies of the effectiveness of this model, we can be clear about the alternatives before us. The risk of failing to improve leadership evaluation is grave; the risk of providing a fair, specific, and constructive leadership evaluation system is minimal. Multidimensional Leadership Assessment is worthy of serious consideration and critical evaluation.

References

Danielson, C. (2002). *Teaching evaluation*. Alexandria, VA: ASCD.

Darling-Hammond, L., & Sykes, G. (Eds.). (1999). *Teaching as the learning profession: Handbook of policy and practice*. San Francisco: Jossey-Bass.

Dipaola, M., & Tschannen-Moran, M. (2003). The principalship at a crossroads: A study of the conditions and concerns of principals. *NASSP Bulletin, 87*(16), 43–65.

Interstate School Leaders Licensure Consortium. (1996). *Standards for school leaders*. Washington, DC: Council of Chief State School Officers.

Junger, S. (1998). *The perfect storm*. New York: Harper.

Reeves, D. B. (2002a). *The leader's guide to standards: A blueprint for educational excellence and equity*. San Francisco: Jossey-Bass.

Reeves, D. B. (2002b). *The daily disciplines of leadership: How to improve student achievement, staff morale, and personal organization*. San Francisco: Jossey-Bass.

Douglas B. Reeves is Founder of the Center for Performance Assessment; dreeves@makingstandardswork.com. His most recent book is *Accountability for Learning: How Teachers and School Leaders Can Take Charge* (ASCD, 2004).

The Case Against the Zero

Even those who subscribe to the "punishment" theory of grading might want to reconsider the way they use zeros, Mr. Reeves suggests.

BY DOUGLAS B. REEVES

THIS IS not a trick question. If you are using a grading scale in which the numbers 4, 3, 2, 1, and 0 correspond to grades of A, B, C, D, and F, then what number is awarded to a student who fails to turn in an assignment? If you responded with a unanimous chorus of "zero," then you may have a great deal of company. There might be a few people who are familiar with the research that asserts that grading as punishment is an ineffective strategy,[1] but many of us curmudgeons want to give the miscreants who failed to complete our assignments the punishment that they richly deserve. No work, no credit — end of story.

Groups as diverse as the New York State United Teachers and the Thomas Fordham Foundation rally around this position.[2] Let us, for the sake of argument, accept the point. With the grading system described above, the failure to turn in work would receive a zero. The four-point scale is a rational system, as the increment between each letter grade is proportionate to the increment between each numerical grade — one point.

But the common use of the zero today is based not on a four-point scale but on a 100-point scale. This defies logic and mathematical accuracy. On a 100-point scale, the interval between numerical and letter grades is typically 10 points, with the break points at 90, 80, 70, and so on. But when the grade of zero is applied to a 100-point scale, the interval between the D and F is not 10 points but 60 points. Most state standards in mathematics require that fifth-grade students un-

DOUGLAS B. REEVES is the chairman and founder of the Center for Performance Assessment, Boston, Mass. His most recent publications are Assessing Educational Leaders (Corwin Press, 2004) and Accountability for Learning (Association for Supervision and Curriculum Development, 2004).

derstand the principles of ratios — for example, A is to B as 4 is to 3; D is to F as 1 is to zero. Yet the persistence of the zero on a 100-point scale indicates that many people with advanced degrees, including those with more background in mathematics than the typical teacher, have not applied the ratio standard to their own professional practices. To insist on the use of a zero on a 100-point scale is to assert that work that is not turned in deserves a penalty that is many times more severe than that assessed for work that is done wretchedly and is worth a D. Readers were asked earlier how many points would be awarded to a student who failed to turn in work on a grading scale of 4, 3, 2, 1, 0, but I'll bet not a single person arrived at the answer "minus 6." Yet that is precisely the logic that is employed when the zero is awarded on a 100-point scale.

There are two issues at hand. The first, and most important, is to determine the appropriate consequence for students who fail to complete an assignment. The most common answer is to punish these students. Evidence to the contrary notwithstanding, there is an almost fanatical belief that punishment through grades will motivate students. In contrast, there are at least a few educators experimenting with the notion that the appropriate consequence for failing to complete an assignment is to require the student to complete the assignment. That is, students lose privileges — free time and unstructured class or study-hall time — and are required to complete the assignment. The price of freedom is proficiency, and students are motivated not by threats of failure but by the opportunity to earn greater freedom and discretion by completing work accurately and on time. I know my colleagues well enough to understand that this argument will not persuade many of them. Rewards and punishments are part of the psyche of schools, particularly at the secondary level.

But if I concede this first point, the second issue is much more straightforward. Even if we want to punish the little miscreants who fail to complete our assignments — and I admit that on more than one occasion with both my students and my own children, my emotions have run in that direction — then what is the fair, appropriate, and mathematically accurate punishment? However vengeful I may feel on my worst days, I'm fairly certain that the appropriate punishment is not the electric chair. Even if I were to engage in a typically fact-free debate in which my personal preference for punishment were elevated above efficacy, I would nevertheless be forced to admit that giving a zero on a 100-

point scale for missing work is a mathematical inaccuracy.

If I were using a four-point grading system, I could give a zero. If I am using a 100-point system, however, then the lowest possible grade is the numerical value of a D, minus the same interval that separates every other grade. In the example in which the interval between grades is 10 points and the value of D is 60, then the mathematically accurate value of an F is 50 points. This is not — contrary to popular mythology — "giving" students 50 points; rather, it is awarding a punishment that fits the crime. The students failed to turn in an assignment, so they receive a failing grade. They are not sent to a Siberian labor camp.

There is, of course, an important difference. Sentences at Siberian labor camps ultimately come to an end, while grades of zero on a 100-point scale last forever. Just two or three zeros are sufficient to cause failure for an entire semester, and just a few course failures can lead a student to drop out of high school, incurring a lifetime of personal and social consequences.

This issue is as emotional as anything I have encountered since the phonics versus whole language debate. Scholars regress to the persuasive tactics of professional wrestlers (no offense intended to wrestlers — this article will generate enough hate mail as it is), and research and logic are subordinated to vengeance masquerading as high standards. Because the emotional attachment to the zero is so strong, I have given up advocating that 50 points should represent the lowest grade. What I do think we can do to preserve some level of sanity in our grading system is to return to a four-point system. A's no longer equal 100 points, but four points. If there is a need for greater specificity, then we can choose an infinite number of digits to the right of the decimal point and thus differentiate between the 3.449 and 3.448 to our heart's content. But at the end of the day in such a system, the F is a zero — one point below the D. It is fair, accurate, and, some people may believe, motivational. But at least the zero on a four-point scale is not the mathematical travesty that it is when applied to a 100-point system.

1. Thomas R. Guskey and Jane M. Bailey, *Developing Grading and Reporting Systems for Student Learning* (Thousand Oaks, Calif.: Corwin Press, 2001).

2. Clarisse Butler, "Are Students Getting a Free Ride?," *New York Teacher*, 2 June 2004, available at www.nysut.org/newyorkteacher/2003-2004/040602grading.html; and Thomas B. Fordham Foundation, "Minimum Grades, Minimum Motivation," *The Education Gadfly*, 3 June 2004, available at www.edexcellence.net/foundation/gadfly/issue.cfm?id=151#1850.

File Name and Bibliographic Information

k0412ree.pdf

Douglas B. Reeves, "The Case Against the Zero," *Phi Delta Kappan,* **Vol. 86, No. 4, December 2004, pp. 324-325.**

EDUCATION WEEK

Published: June 6, 2001

COMMENTARY

If You Hate Standards, Learn To Love The Bell Curve

By Douglas B. Reeves

The politically correct blood sport among educational commentators these days is the jeremiad against the evils of academic standards and testing. High expectations and, even worse, testing to ensure that those expectations have been met is, in the accepted creed of the faculty lounge and parent-teacher- organization meeting, the devil's own instrument. As everyone knows, one must "teach to the test" and thus engage in low-level "drill and kill" in order for students to succeed on these mindless examinations.

> Despite the obvious flaws of an educational system based upon academic standards, it is far superior to the available alternatives.

What everyone knows is, of course, wrong. Winston Churchill said of democracy that it is "the worst of all political systems—except for all the others." So it is with standards. Despite the obvious flaws of an educational system based upon academic standards, it is far superior to the available alternatives.

Few analysts have considered the fundamental question: If standards and testing disappeared tomorrow, what would be the alternative? To hear the critics of standards and tests, the answer would be educational paradise. Such an assumption rests upon the faith that, absent standards and testing, every classroom would offer expectations that were clear, rigorous, and objective. Success in one grade would be related to success in the next grade, because communication and coordination among teachers and different grade levels would be flawless. Without external standards and expectations, the testing conducted by teachers would be inherently fair because it would be based upon the achievement of an objective result rather than comparison of one student to the other.

> If standards and testing disappeared tomorrow, what would be the alternative?

Paradise, alas, eludes us. The alternative to standards and testing is not educational nirvana, but a return to the bell curve and its twin, mystery grading. Without objective standards, the basis of comparison for students is not the relationship of student work to an objective standard, but the comparison of one student to the other. A few students will succeed, a few will be tossed onto the academic scrap heap, and the vast majority will be "normal" and thus fit the distribution that characterized the eugenics movement and the educational establishment for decades. The abandonment of standards leads us to the era in which teachers identified bluebirds, robins, and blackbirds, the choices of color hardly an accident. The abandonment of testing embraces the world in which we have not perfection, but grading as the mysterious determination of the teacher. What parent has not endured the following conversation? "What did you do in school today?" *Nothin'.* "Why did you get that grade?" *I dunno.* From the mouths of babes shall come the truth. Only clear standards and consistent assessments offer a coherent response to these entirely reasonable questions.

The fundamental flaw in the reasoning of the critics of standards and testing is this: However

47

much they decry the evils of standards and tests, the alternative is worse. The alternative to standards is the bell curve, in which teachers have for decades compared student performance not to an objective standard, but to that of other students. This has provided the worst of both worlds: Proficient students have been labeled as failures because they failed to achieve scores higher than their more proficient colleagues'; nonproficient students have been complacent because they were able to beat their less proficient peers. If standards and state tests were eliminated tomorrow, every school in the nation would be left with the absurdity of students who cannot read, write, or compute at levels appropriate for their grades feeling full of false self-esteem because they scored "above average" when compared with their even less adequate peers. Worse yet, students making great progress and at last performing at a proficient level would be regarded as inadequate because their proficiency was a percentile below another child's.

The argument against academic standards and rigorous tests rests upon a syllogism that is honored more in the passion with which it is expressed than the evidence supporting it. The syllogism asserts first that academic standards are narrow in scope and inherently focused on "mere facts" rather than deep thinking and analysis. Second, the only way that one can succeed on standards- based tests is to engage in mindless test preparation and the inherently evil "drill and kill" exercises of the medieval classroom. Third, the good teacher who insists upon rigorous analysis, reasoning, thinking, and writing would produce students doomed to failure on standardized tests.

Reduced to its essence, this syllogism asserts that bad teaching yields good test scores, and that good teaching yields bad test scores. Not only does the evidence fail to support this common assertion, the data on the subject lead to precisely the opposite conclusion. Research from the Center for Performance Assessment, multiple other sources, and that rarely considered factor, common sense, reveals that teachers who focus on analysis, reasoning, thinking, and particularly writing not only have challenging classrooms and literate students, but also produce pupils with higher scores on state and district tests.

> The alternative to standards and testing is not educational nirvana, but a return to the bell curve and its twin, mystery grading.

Neither of the prevailing political extremes, however, seems interested in the evidence. Some protestors oppose standards and testing, firm in the conviction that the bell curve is true and that without social intervention, poor and minority children cannot succeed. Such an inherently racist premise does scant justice to the good intentions of those who advocate this position. Such a view also ignores the mountain of evidence that demographics are not destiny.

The far right opposes standards and testing, convinced that somehow the political aphrodisiac of "local control" will shrink if every student is required to read, write, and compute. The appeal for local control creates a curious alliance between the far right and militant test protestors, both of whom apparently believe that their rights are threatened if the public learns about the reading levels of schoolchildren. When confronted with the fact that an astonishing number of 8th grade students are unprepared for the literacy and mathematical demands of high school courses, some advocates find it easier to screech an oration on the benefits of local control or the perils of tests than to teach students to read, write, and compute. Neither side appears interested in, much less convinced by, evidence that rigor, analysis, writing, editing, and hard work by students and teachers yield better results than either mindless test prep or endless whining about testing.

Some state academic standards are indeed narrow in scope, while others focus on broad issues of analysis and understanding. But the proposition that this is an irreconcilable paradox is laughable to the workaday teacher, who understands that the concepts of mathematical problem-solving will elude students who cannot add, subtract, multiply, and divide.

Teachers toiling in the vineyard with real students also regard as preposterous the notion that students can apply high-order thinking skills to history, geography, and economics without understanding that the Civil War preceded Vietnam, that the Balkans are not the Baltics, and that there is rarely a singular cause for an historical or economic effect. Regardless of the language of state standards and the contents of state tests, good teachers routinely provide a combination of factual knowledge and analytical understanding.

State tests have similarly inevitable failings. Tests fail to reflect the full scope and complexity of the curriculum of schools. In fact, no test can or should examine every element of every curriculum in every school.

The demands of the legislators and those whom they represent, the parents of today's schoolchildren, are much more modest than a comprehensive evaluation of education.

> No thoughtful advocate of academic standards and rigorous state tests argues that the present state of the art is perfect.

We simply want to know if our kids can read, write, and compute, and we are not willing to concede that such a request constitutes child abuse, despite the histrionic claims of the anti-standards movement.

No thoughtful advocate of academic standards and rigorous state tests argues that the present state of the art is perfect. Standards should be clearer and more rigorous. Tests should be more comprehensive and clearly related to standards. I should be tall and handsome. The remedy for two of these three deficiencies is perseverance, hard work, and collaborative effort by people of goodwill. None of these deficiencies will be remedied by the abandonment of standards or the elimination of the measuring stick.

The clarion call for boycotting tests and abandoning standards recalls the desire of the obese chain-smoker to hide the scales and discard the blood-pressure cuff. When we don't like the results, blame the instruments.

Most parents and many thoughtful but silent educators know that hiding the results will not improve the health of the patient. We can handle the truth, but we have a diminishing tolerance for those who prefer fact- free self-congratulation to turning off the television and video games, opening the backpack, and finishing our collective homework.

Douglas B. Reeves is the president of the Center for Performance Assessment, based in Denver, and the author of 10 books, including, most recently, Crusade in the Classroom: How George W. Bush's Education Reforms Will Affect Your Children, Our Schools (Simon & Schuster, 2001).

Center for Performance Assessment

Clear Answers to Common-Sense Questions about Accountability

by Dr. Douglas Reeves
Thrust for Educational Leadership (March/April 2000)

Parents and policy-makers have a direct agenda when it comes to accountability. Their questions must be addressed with clarity and candor.

George Orwell (1949) helped generations of students learn the meaning of the term "double-speak." Few practitioners of the art of double-speak have studied the concepts as assiduously or applied it with as much vigor as attackers and defenders of public education. Both the strident tone of the debate between these two camps and the obscurity of the arguments employed by the debaters cause many observers to throw up their hands in dismay.

Parents and policy-makers have a more direct agenda. In terms of educational accountability, they express the issues very clearly.

"How's my kid doing?"

"Are the schools succeeding or failing?"

"What works best to help students learn?"

"Do test scores prove the effectiveness of educational programs?"

These are serious questions, and effective accountability systems must address them with clarity and candor. An accountability system that fails to address these common-sense questions does not deserve the support and confidence of citizens or policy-makers.

Simple questions, however, do not necessarily lead to simple (or, more appropriately, simplistic) answers. So it is with educational accountability. These simple questions require complex responses, which I will explore in this article.

How's my kid doing?

A question about the performance of a specific child implies that an effective accountability system will be based on information about individual students rather than groups or classes of students. In other words, an accountability system that can supply information about the progress of a specific child must be based on a series of individual student achievement records. If these records are flawed, then the entire house of accountability is built upon a shaky foundation.

One of the most important developments in educational accountability has been the "value-added" methodology developed by Professor William Sanders and his colleagues at the University of Tennessee (1998). The core of the system is a "student-to-student" comparison. While such a comparison makes common sense, it is rare. By far the majority

50

state and local accountability systems compare one year's class of students to the previou year's class of students - a comparison that involves almost entirely different individual students.

Such group comparisons never address the fundamental question, "How is my kid doin Rather, parents and teachers are given the curious information that their eighth-grade students are better or worse than last year's eighth-grade students. Group comparisons te us nothing about the progress and educational needs of individual students.

This has serious implications for accountability systems that claim to evaluate educatio quality, but ignore information about classroom teaching and learning. An accountability system that contains test scores alone, without the context of additional accountability information about teaching practices and curricula, is incomplete.

A school district that bases its accountability system on test scores alone is analogous a physician who evaluates physical health based upon indicators such as body temperatui or blood pressure, but ignores the other medical indicators that any reasonable physician would regard as essential to a competent diagnosis. In the most bizarre cases, accountab systems not only fail to evaluate substantial parts of the school curriculum, but actually encourage behaviors contrary to those endorsed by the designers of the accountability system (see example).

If parents or stakeholders want an honest and accurate response to the question of "hc my kid doing?" then school systems and districts must use tests or other assessments tha will yield the information needed to answer this question. In other words, children need to I assessed to make sure they meet the standards, and only this information will determine I they're "doing."

The most virulent critic of public education would not attend an athletic event and, desii to know the outcome, settle for a description of the weather and field conditions. "What the heck happened during the GAME?" the critic would demand. In the context of the classroc parents and students must learn about the process and results of teaching and learning.

If the education "game" is to be taken at least as seriously as many people take their weekend athletic events, then it is reasonable to expect that the data used to evaluate the result should be related to the issue at hand: student and school performance. Thus, if we want to know how well students write, then we must ask them to write. If we want to know whether they can use the scientific method, then we should ask them to design an experiment and draw inferences from a set of data. If we wish to know whether students understand mathematics, we should ask them to use mathematics to solve real problems.

Are schools succeeding or failing?
The second question raised in the name of common sense is built upon the first. Schoc succeed only where student succeed. Thus, assessments of school such as accreditation typical accountability reports, are only as effective as their evaluation of students. Accountability systems that look only at process and effort will reward a fixation on meetin plans and strategies while ignoring results. Every initiative, including those that I have advocated - high standards, effective assessments and rigorous accountability - is only effective when it is built upon a foundation that soundly evaluates student achievement.

Mike Schmoker (1999, 1999a) is a leading advocate of the focus on results. He has endured heaps of abuse from those who find process a perfectly suitable substitute for student achievement. While a focus on results is important, the "results" that effective accountability systems must consider extend far beyond test scores alone.

This is not merely a debate over "process vs. results" In fact, a comprehensive accountability system must focus both on measurable elements of the process (specific

instructional, assessment and engagement strategies), and on results (indicators of studei achievement). Only with such a comprehensive view can we gain some insight into what tl adults in the system can do to influence results achieved by the students.

Although the evaluation of students is the foundation of a sound accountability system, effective accountability system must base its examination of student achievement on more than test scores. Indeed, the fallacy of reporting school-wide success of failure based on single test scores has been widely documented (Bracey 1999). Teaching and learning are multiple interactive processes, the results of which are much too complex to be captured t single score.

Accountability systems that depend solely on test scores offer predetermined results: students who are capable test takers will make a school look like a success; students who are not capable test takers will give their school the label of failure. In neither case do the organization, leadership, teaching and educational practices of the school receive a meaningful evaluation.

Instead, effective evaluation of the success of a school can only be measured with multiple information sources over an extended period of time. There must be several indicators that measure the performance not only of students, but also of the adult decisioi makers. Such an analysis includes a consideration of resources, teaching methods and student support. Without a consideration of all of these factors, we are left with the ludicroi notion of evaluating the performance of an entire institution based on the test scores of on group of eight-year-olds during one stress-filled, springtime afternoon.

Effective, comprehensive accountability systems distinguish between student achievement and school achievement without losing sight of the fact that the latter is integrally related to the former. For example, state and district test scores might provide some indicators of student achievement. But these results are only meaningful indicators (school achievement when they are placed in the context of the specific educational strateç used by the schools. The structure of such a comprehensive system acknowledges both tl distinction and the importance of both questions.

What works best to help students learn?
The third common-sense question addresses the heart of an effective accountability system. The question of program effectiveness is far more complex than the recitation of t scores representative of most accountability systems. Such a "box score" approach is the educational equivalent of giving the ranking of teams at the end of the season without shedding any light on the strategies that led the teams to succeed or fail.

For the disinterested observer who wishes only a cursory overview, the final rankings n suffice. But those interested in the game would want much more. People whose futures depended upon the success of the team would demand a continuous analysis, not only of scores, but also of the individual elements of strategies and programs that lead to success

In the context of school accountability, stakeholders need to know which programs succeeded and which ones failed. In a field littered with "reforms" and "new ideas," some rational method of evaluation other than popularity, enthusiasm or cost must be used. An accountability system that shows policy-makers how intervention strategies correlate with student results can go a long way toward providing such essential program evaluation information.

Some of the best practitioners of such systematic evaluation are Robert Slavin and his colleagues at John Hopkins University. In their recent book, "Show Me the Evidence!" (19(they provide examples of how accountability information can be used to monitor program effectiveness. Perhaps the most astonishing aspect of the book is how few programs have been subject to long-term, systematic accountability and analysis.

In sports, media commentators are expected to report not only scores, but they must al demonstrate a thorough understanding of the nuances of the game. They routinely debate the meaning of various statistics. The television commentator or sports journalist who simp writes, "Those with winning records are better teams than those with losing records" woulc soon be out of a job.

When it comes to sports, we expect the best commentators to look beyond the data an provide insights based on observation, description and qualitative understandings that ext beyond numerical explanations. Though the athletic analogy may seem clumsy, I will celebrate the day when analysis of educational accountability data is taken as seriously by the media and the public as analysis of last weekend's sports games. When that day com we all may finally know what works best to help students learn.

A comprehensive evaluation of what works best to help students learn must go far bey mere test scores and include analyses of both how the data reported should be interpreted as well as information about the context in which these results were derived.

Balancing qualitative and quantitative information
In educational accountability, numbers are an important part of the story, but they tell o part of the story. The qualitative dimension of accountability - descriptions, narratives and observations about culture and climate - creates a lens through which the quantitative data must be viewed.

What is the qualitative context of the quantitative data? What are the successes, failure tragedies, and triumphs of this school that help to explain the story behind the numbers? While no principal in America needs one more report to write, a one-page synopsis of a school's qualitative dimension would add greater context to its test scores. Without such a qualitative context, we are left with the sterility of data that, even when presented with abundant statistical complexity, can leave us wondering, "What really happened in that school?" Without the qualitative dimension, our understanding of the "score" is limited, incomplete and possibly inaccurate.

What sort of qualitative information should be included? Information about the school climate and environment, the triumphs and tragedies of the school year, and descriptions any significant changes in programs, personnel or performance can all be expressed in narrative form. Such narrative information is not a substitute for quantitative data, but rathe gives citizens and policy-makers context in which to interpret numerical results.

Meeting the test of common sense
An effective accountability system must answer at least four common-sense questions one about individual student achievement, a second about school performance, a third ab ways to help students learn, and a fourth about determining educational effectiveness.

In order to provide useful information about student achievement, an accountability system must be based on clear standards that have been communicated to students, parents, teachers, and other district stakeholders. Both quantitative and qualitative indicato that measure whether or not these standards have been met must become integral parts c the accountability system.

School performance must be based on much more than test scores. Though it is likely include test data, school performance information must also include how those numbers should be interpreted and the context from which test scores arose. This approach to comprehensive accountability is, to be sure, more challenging than simplistic headlines. It however, the only approach that meets the simple test of common sense. [References]

Accountability is More Than Test Scores

by Dr. Douglas Reeves
Advanced Learning Press (1998)

Test scores represent only one piece of the accountability data. These test scores should interpreted in the context of other information about what schools are doing.

The Accountability Imperative

Accountability is an imperative facing every school administrator and board member. Unfortunately, what passes for "accountability" is frequently no more than a recitation of te scores. Such test scores may or may not reflect the accomplishments of students, teacher and leaders in a school system. Unfortunately, these numbers often influence the jobs of superintendents and principals, property values in a neighborhood, and the confidence of public in our educational system.

Excellence and Equity

Some districts view accountability solely as a compilation of test scores. This can become thinly veiled mechanism to threaten administrators and teachers. Other school districts ha' found a better way. These districts have transformed accountability into a constructive strategy for informing the public and reforming schools. They have focused attention not o on test scores, but also on the "antecedents of excellence." These antecedents of excelle are those activities in schools that lead to the twin goals – excellence and equity.

Five Key Principles

There are five key principles for effective accountability systems: congruence, specificity, relevance, respect for diversity, and continuous improvement. Although many accountabili systems incorporate some of these principles, it is rare to find one that includes all of them

1. Congruence

Standards, assessments, and evaluations of teachers and administrators must be congrue with the accountability system. Many teachers are frustrated when school administrators establish academic standards and then hold teachers accountable for test scores that are only marginally related to those standards. The same frustration afflicts superintendents when a district or state accountability system is announced with considerable fanfare, but evaluations of administrators and teachers, as well as the assessments of students, fail to match the goals of the accountability system. Accountability is nothing more than a paper tiger without congruence.

2. Specificity

Educational reformers are sometimes better able to articulate grand goals than to identify specific expectations. "Improve reading achievement" is a wonderful goal, but it gives little the way of specific advice for teachers and principals. A better goal is to "increase by ten percent the proportion of students who receive a score of "proficient" or better on the distri reading assessment." Teachers and administrators know precisely what is expected of the and how to prepare students for this challenge.

3. Relevance

Principals and teachers need accountability goals that are directly related to their individue schools. Test scores are related to effects. Meaningful accountability systems have goals are relevant to causes. In the Milwaukee Public Schools, for example, a nationally recogni effort was undertaken to improve math performance by students. Schools were not only recognized for the achievement of high math scores. They were also recognized for achieving the antecedents of excellence – improving calculator proficiency, establishing "math buddies" programs, increasing parent tutor training, and a host of other accountabili indicators – that were directly relevant to the needs of the schools and the overall goal of improved math scores.

4. Respect for Diversity

Effective districts have high expectations for all students — rich and poor, minority and majority. All students are expected to achieve at high levels. Effective accountability syste should provide for some indicators that are uniform throughout the district and others that unique to each school. A system that combines "system-wide" indicators (such as achievement in math, reading, and science) with "school-based" indicators, can provide th critical balance. In Milwaukee, the accountability system included six system-wide measur and five school-based measures. The school-based measures were chosen by teams of administrators, teachers, and parents. The diversity of the district was reflected in the diversity of the school-based goals. They range from an emphasis on discipline to an emphasis on participation in Advanced Placement calculus classes. All of these school-based indicators had one thing in common: they addressed the individual needs of that school and represented a specific mechanism for helping that school achieve the system-wide goals of higher student achievement.

5. Continuous Improvement

The fundamental idea is this: accountability systems must themselves be accountable. School leaders want to ensure that the focus remains on high student achievement, the expectations of rigor and challenge are consistent, and there is congruence between the accountability system and the expectations of the community. Thus the accountability syst cannot be regarded as holy writ, exempt from continuous review and improvement. Stakeholders from throughout the community should regularly review the accountability indicators to ensure these measures continue to meet all of the previous four principles an serve the needs of the community.

The Typical System

Many superintendents will insist, "We already have an accountability system." This is certainly true. Unfortunately, many of these systems share the same characteristics that inhibit their effectiveness.

Erroneous Presumptions

First, there is an erroneous presumption that accountability is the same as a compilation o test scores. We would be shocked if our physicians recommended major surgery based or our body temperature, weight, and a few other scattered indicators. Yet we recommend

major surgery for our school systems based on test scores of dubious quality and relevanc
I am not an apologist for low test scores. However, I do maintain that these scores represe
only one piece of the accountability data. Therefore, these scores should be interpreted in
context of other information about what schools are doing.

No Comparability

The second characteristic of typical accountability systems is a lack of comparability. The
typical "spring-to-spring" comparisons assume a homogeneity between grades that does r
exist in many school districts, particularly those with high mobility rates. Even teachers in
more stable rural areas are well aware that testing students in a single grade and compari
the fourth graders of 1995 to the fourth graders of 1996 is not a reflection of academic
achievement, but rather a reflection of the characteristics of very different students. In
Milwaukee, schools were given the choice of making "spring-to-spring" comparisons or "fa
to-spring" comparisons. In schools with high mobility rates, the latter provided a much mor
accurate indicator of school improvement.

Quantitative Data

The third characteristic of typical accountability systems is that they are limited to quantitat
data. I have taught multivariate statistics to graduate students and would summarize my
knowledge of statistical analysis as follows: numbers alone never tell the whole story. A
narrative description of the challenges and opportunities from each school would create a
lens through which quantitative data can be viewed. Without this lens of descriptive
information, the numbers are sterile and devoid of context.

Fixation on Thresholds

The fourth characteristic of typical accountability systems is that they are fixated on
thresholds and not improvements. Schools are often given pejorative labels such as "Schc
In Need of Improvement" or its unfortunate acronym "SIN Schools." In addition to feeding ·
dogmatism of those who enjoy attacking public education, these labels are imprecise and
inaccurate. They imply a binary system in which schools are either "OK" or "Not OK" and f
to investigate the extent to which schools are making progress and addressing the
antecedents of excellence. Two schools might have identically low test scores. However, (
that is making substantial progress on the antecedents of excellence (reported in its
accountability system) has a significantly different profile than the school that cannot rend(
such as report. Policy makers and the public ought to have sufficient information to
distinguish between these two schools.

*A comprehensive accountability system
is more rigorous than a list of test scores*

A Better Way

The Comprehensive Accountability System

There is a constructive alternative to the typical accountability system. The Comprehensiv
Accountability System, such as that implemented by the Milwaukee Public Schools, includ
three separate tiers.

First Tier

The first tier consists of six "system-wide" indicators such as math, reading, writing, scienc
grades, and attendance. For each of these six indicators, schools can earn from zero to te

points based on their achievement of the goals established by the Board of School Directc
These high goals are the same for every school in the district.

Second Tier

The second tier consists of five "school-based" indicators. In these indicators, each school
has identified its own unique antecedents of excellence. These indicators reflect the divers
of the district. Some schools choose to focus on technology, staff development, discipline,
parental involvement. All schools, however, choose these indicators so they can improve
their performance on the system-wide indicators. For each of these five indicators, schools
can earn from zero to ten points based on their improvement from the previous year.

Third Tier

The third tier is a narrative description of the challenges and opportunities facing the scho(
This page-long narrative allows the reader to better understand the test scores and other
data that are contained in the first and second tier. The Comprehensive Accountability
System creates balanced incentives, encouraging all schools to meet uniformly high
standards while at the same time allowing each individual school to pursue goals that are
uniquely relevant to its needs.

Political Realities Must Be Confronted

There are clearly some political realities that must be confronted by districts attempting to
move beyond the "report card" (or a list of test scores) to an improved Comprehensive
Accountability System. These realities can be anticipated and the objections they imply ca
be answered.

Rigor and Accountability

The first and most important reality is that the public demands rigor and accountability. Th(
Comprehensive Accountability System is more rigorous than a list of test scores. It holds
schools accountable not only for test results but also for specific activities that are necess;
in order to achieve higher scores.

Consequences and Incentives

The second reality is that the public demands consequences and incentives associated wi
school accountability. A Comprehensive Accountability System, fully congruent with
administrator evaluations, will provide this link between public accountability and personal
responsibility. On the other hand, it also ensures fairness to superintendents and principal
by preventing policy makers from enunciating an accountability policy and then issuing
evaluations that are only remotely related to the accountability goals.

Faster

The third reality is that a Comprehensive Accountability Plan provides faster reporting and
analysis of information. The school-based indicators are half the system. These can be
gathered and analyzed during the school year. Rather than waiting months for test scores.
schools can make "mid-course corrections" that are at the heart of an effective data-driver
decision-making process.

Conclusion

Demands for accountability are here to stay. Superintendents and school board members

57

must choose whether to allow accountability systems to be a "gotcha!" used by public education critics or to create accountability systems that are rigorous, relevant, fair, and meaningful. The Comprehensive Accountability System is a step in the right direction.

NOTE: *This article is based on a presentation to the national convention of the American Association of School Administrators (AASA) in 1997. The author gratefully acknowledges the assistance of Mr. Robert Jasna, Supt.; Mrs. Barbara Horton, Deputy Supt.; Dr. Deanne Housfeld, Assist. Deputy Supt.; and the principals, teachers, parents, and community who served on the Accountability Task Force in Milwaukee.*

Standards Are Not Enough: Essential Transformations for School Success

Douglas B. Reeves

This article discusses five transformations necessary to bring standards from theory into reality. These changes, though difficult to elicit, will have a positive and productive effect when made collaboratively and thoughtfully. The concept of power standards, a subset of standards that involves thoughtful focus, is proposed to ensure successful educational practice and improved test performance.

Almost every public school district in the nation has academic standards. Although standards alone are clearly an insufficient instrument for the improvement of student achievement, the essence of standards—the clear articulation of what students should know and be able to do—forms the basis for the essential transformations necessary for school success. In this article I discuss five essential transformations necessary to bring standards from theory to reality:

1. A change in instructional strategies from the extremes of mindless test preparation and unsupervised anarchy to a consistent requirement for thinking, reasoning, and communication proficiency for all students.
2. A change from the guesswork and alchemy surrounding letter grades to the use of comprehensive diagnostic assessments before and during high school.
3. A change in curriculum from frantic coverage of every standard to the use of "power standards."
4. A change from assigning teachers to buildings and courses on the basis of seniority to a more equitable assignment policy that is blind to the economic status and skin color of students.
5. A change from high-stakes tests as the sole criterion for critical decisions to the use of a body of evidence for high-stakes decisions.

These changes strike at the heart of some of the most hallowed traditions of secondary education. But I will be clear: The choice is not whether these changes will occur, but whether they will be made collaboratively and thoughtfully or through a haphazard combination of politics, state intervention, and litigation. For those who favor the former process of change, the following ideas may have some value (see figure 1).

Transformation 1. From Test Prep to Thinking, Reasoning, and Writing
Many individuals believe that the appetite of politicians for higher test scores can be satisfied only when teachers engage in mindless test preparation at the expense of thinking, reasoning, and writing. Consider the words of standards critic Alfie Kohn (1999):

Not long ago, I am told, a widely respected middle school teacher in Wisconsin, famous for helping students design their own innovative learning projects, stood up at a community meeting and announced that he "used to be" a good teacher. These days, he explained, he just handed out textbooks and quizzed his students on what they had memorized. The reason was very simple. He and his colleagues were increasingly being held accountable for raising test scores. The kind of wide-ranging and enthusiastic exploration of ideas that once characterized his classroom could no longer survive when the emphasis was on preparing students to take a standardized examination. Because the purveyors of Tougher Standards had won, the students had lost. (73)

Many educators share experiences similar to those addressed by Kohn, and this anecdotal evidence suggests that some administrators and teachers embrace a regime of test preparation through the repetitious use of mock test items. This leads to the widespread conclusion that the very existence of standards and associated tests leads to bad decisions by administrators about teaching strategies. At one end of the rhetorical continuum are the fearful administrators demanding a curriculum laden with test preparation activities. At the other end are those who contend that no supervision is appropriate and echo the mantra, "Just leave me alone and let me teach!" Consider the words of Susan Ohanian (1999):

What I do know is that across the land teachers are bowing to the pressure of corporate-political-infotainment pronouncements: They are eliminating recess and putting away the building blocks, the tempera paints, and the picture books that don't introduce phonemes in the sequence chosen by the publishing conglomerate so venerated in Texas and California. They are bringing out the skill drill worksheets that will get every kid in America learning the schwa on schedule. (19)

These comments beg the question, Does test preparation (along with the subsequent abandonment of thinking, reasoning, and communication skills) improve test scores? If the answer to this question is yes, then there ought to be data to support such a notion. Instead, the key to better multiple choice scores lies in a focused, multidisciplinary requirement for students to think, reason, and write in a clear, accurate, and persuasive manner.

If educators could step back from the rhetorical wars surrounding high-stakes testing, it should become obvious that thinking rather than memorization is the key to better test results. An examination of case-study data at the classroom, school, district, and state level indicates that when students write more frequently and when they score higher on writing performance assessments, their scores on multiple-choice tests improve. The association between writing and multiple-choice scores ranges from .7 to .9 (see figure 2). Other case studies in Riverview Gardens, Mo., Milwaukee, Wisc., and Orange County, Calif., support this quantitative evidence. These case studies in a variety of school systems including those with high levels of poor and minority students confirm a strong relationship between an increased frequency of writing performance assessments

and improved test scores. More writing in the classroom may, indeed, take time away from test preparation, but it does not hinder test scores.

Does an emphasis on more writing and its attendant time demands diminish other subjects in the curriculum? The evidence from Riverview Gardens, Mo., school district suggests the opposite. The Write Focus program comprises collaboratively scored writing assessments in every class, including physical education, music, and all other disciplines, on a monthly basis. Using a common rubric or scoring guide, the district tripled the number of students who were proficient or better in writing. According to an editorial in the St. Louis Post-Dispatch (2000), this "diversion" of time and effort from other subjects had an interesting consequence: dramatic increases in state test scores in science and social studies. Similar associations between strong writing programs and cross-academic improvements can be found in Orange County, Calif. (Groves 2000); Milwaukee, Wisc.; and Indianapolis, Ind. In Wayne Township in Indiana, for example, there is a common writing scoring guide for grades 7 through 12 so that students from science to sociology and mathematics to music receive consistent messages on the importance of expressing their reasoning and thinking in appropriate written form. These model districts make it clear that reasoning, thinking, and communications—not mindless test drills—are the keys to improved student achievement.

The administrators who demand test preparation are wrong; the standards critics who assert that only terrible teaching practices will lead to higher multiple-choice test scores are also wrong. The positive impact of writing, thinking, and reasoning on test scores is no surprise to secondary school educators who have successfully prepared students for Advanced Placement (AP) and International Baccalaureate (IB) exams for years. The rigor of those classes, with their emphasis on effective communication and thinking and a fundamental understanding that one cannot possibly memorize every conceivable test item, is good educational practice for all students, not merely those with the designation of "advanced."

Thus the first essential transformation for successful schools is to challenge both extremes in this acrimonious and unenlightening debate. Micromanagement of teachers with mindless test prep is not effective and demoralizes the best teachers. However, abandoning students to the whim of teachers who do not like standards but are also unwilling to engage in systematic thinking, reasoning, and writing is neither effective nor fair to students. The plain fact is that competent teachers require frequent thinking, reasoning, and writing from students. The most effective writing programs include collaborative scoring by teachers who use a consistent scoring guide across all disciplines. Administrators who facilitate this collaboration and require more writing rather than more test preparation are likely to be rewarded in their efforts with more motivated teachers and more successful students.

Critics of standards and testing will respond that this evidence is irrelevant if administrators do not believe it; fearful administrators will reject thinking, writing, and reasoning in favor of test drill, and thus the problem remains. Their point is well taken; it is not leadership demands per se that are wrong but leadership demands inconsistent with

the evidence that are unproductive. Schools that demand more writing will produce more students with improved skills in thinking, reasoning, and communication. These schools will also benefit from higher test scores.

Transformation 2. From Mystery Grades to Assessments that Are Accurate, Fair, and Educative

Too often, educational tests, grades, and report cards are treated by teachers and parents as autopsies when they should be viewed as physicals. That is, assessments are often perceived as only summative in nature when they should be formative—helping teachers "diagnose" student difficulties and gain insight into how to improve students' achievement. Wiggins (1988) deems these incisive evaluations "educative assessments": they provide feedback that promotes learning rather than letters that signify a mere evaluation. Educators should consider four critical practices to move toward more educative assessment.

First, educators must abandon the average, or arithmetic mean, as the predominant measurement of student achievement. Often, proficient students receive low grades because of failures early in the semester, whereas students in the same class who lack proficiency in the subject but offer consistent performance and good attitudes receive the same grade. The public would never accept such inaccuracy in something it values, such as sporting contests. In that domain, the public insists that the scores reflect the proficiency of the players, and rule changes are sometimes implemented to elevate the relative importance of skill over chance. Yet the public routinely tolerates inconsistent and inaccurate grading systems in those areas of less concern, such as the academic performance of students. Although it is true that students acquire and process information at different rates, it is not the pace of learning that standards-based systems should evaluate. Perhaps a state or district will promulgate the standard that one must "learn algebra quickly." Until that time, however, accurate assessment will focus on the fact that a student is proficient in a subject and not on the pace at which proficiency is acquired.

The practical impact of abandoning the average goes beyond the method of calculating grades. The acknowledgement of variable paces of learning also requires acceptance of variable schedules. In a standards-based school, one student may require four semesters to become proficient in algebra whereas another student requires one semester. If the academic standards refer to achievement rather than pace, then it is the schedule that must be the variable, not the expectations of the teacher or the achievement of the student. This renders obsolete the common requirement that students have three years of math or two years of science or four years of English. In fact, some students need 16 semesters—two periods each day for four years—of English language arts to achieve the state literacy standards. Other students might need 12 semesters of mathematics to acquire proficiency in algebra and geometry. A growing number of secondary schools are making such adjustments in their schedules, making time the variable and maintaining high expectations for all students.

The second change in assessment practice educators must make is a rejection of is the notion that every review of student work results in an evaluation. The assumption

underlying this time-honored practice is that unless work is graded, the students will be unmotivated. This assumption is a splendid one from the point of view of teachers and administrators who, college degrees in hand, think that a D (or, given grade inflation, a B minus) is a terrible grade. Students disagree; they routinely choose a lower grade over the bleaker prospect of more work. In effective schools discussed previously, one of the most consistent practices of successful teachers is the provision of multiple opportunities to learn (Reeves 2000). This does not represent sympathy or low standards, but instead signals a fundamental change in the reaction to unacceptable student work. The consequence for a student who fails to meet a standard is not a low grade but rather the opportunity—indeed, the requirement—to resubmit his or her work. The positive results of this policy include increased respect for teacher feedback and improved learning by students. Such results are only possible, however, when the assessment is not a one-shot affair but rather a process in which students have the opportunity to take teacher feedback, reflect on it, and use it to improve performance. When a final exam is administered on Wednesday and grades are due on Friday, the unspoken message is that teacher feedback is worthless and student efforts to consider teacher feedback and use it to improve academic performance are fruitless.

Third, educators must relate numerical and letter grades with accuracy. Thanks to the predominant use of computer-based grade reports in many secondary schools, teachers are required to use points, typically on a scale of 0 to 100, for student assignments. Teachers then create a system of linking letter grades to these numerical scales, such as A = 93 to 100; B = 85 to 92; C = 77 to 84; D = 69 to 76; and F = less than 69. Although the specifics vary among secondary schools, there is a striking consistency: the interval between grades A through D is markedly different from the interval between grades D and F. In the preceding example, the interval between grades A through D is 7 points, yet the potential interval from grades D to F is 69 points. The result is one with which all students who have ever missed an assignment are familiar: the 0 grade has a disproportionate impact on the average grade. If the teacher had assigned values of 4, 3, 2, 1, and 0 for grades of A, B, C, D, and F, then the impact of a single failure would be mathematically equivalent to a one-grade difference. But when a failure receives a grade of 0 in a 100-point system, the impact of the failure is many times greater than the difference of one grade.

When this inconsistency is pointed out to teachers and administrators, they frequently offer excuses about the manner in which the computer is programmed and the inability of teachers to do much about it. The solution is clear and straightforward: If educators must use a numerical scale, then the lowest possible number on the scale should be the one grade value lower than a D. Using the preceding scale as an example, if the lowest D is a 69 and the interval between grades is 7 points, then the value of a failure should be 69 minus 7, or 62. It certainly should not be 0. This has nothing to do with sympathy or justification for students who fail to meet the standards; it is simply a matter of mathematical accuracy in grading. Without such a commitment to accuracy, a grading system lacks credibility.

Fourth, educators must assume the responsibility to describe completely the grades they assign. This requires the use of a Standards Achievement Report (Reeves 1997) similar to that routinely used in elementary schools. At the secondary level, teachers often raise the objection that an elaboration on the letter grade takes too much time, given a student population of 150 or more for each teacher. In fact, the grade books of secondary educators are full of details, representing thousands of individual marks and many hours of work by the teachers. Unfortunately, these details remain unknown to all but a few students and parents. Those few who discover the labors of the teachers are, unfortunately, the students who challenge a grade and discover that the teacher has documented, with care and precision, the justification for the evaluation. Thus the explanations for letter grades already exist in exquisite detail. The simple change from a one-line-per-student grade book to a one-page-per-student Standards Achievement Report would require no additional effort by the teacher but would provide vastly improved information for students and parents. Figure 3 provides an example of such a report. This example demonstrates that not every assignment encompasses every standard, but every assignment addresses several of the most important standards for this class. Moreover, the teacher has clearly identified that there are both academic and behavioral standards for students. This particular report shown is the profile of a student who is proficient or exemplary in every academic standard, but who displays a rather consistent inability to turn work in on time. The feedback provided by this report is far more revealing than the letter grade.

Transformation 3. From Coverage to Power Standards
The notion of coverage of standards can be a troubling one for teachers. In the casual conversation of the faculty lounge, a teacher may comment that, "Hey, if those kids don't get the volume of a prism, it's not my fault. I can document that I covered it on March 23rd at 10:40 a.m. It's in the lesson plan!" The imperative of coverage of standards is largely the result of the displacement of rigor with girth by the various state academic standard-setting authorities. As Marzano, Kendall, and Cicchinelli (1998) note, the reach of state standards vastly exceeds their grasp. This leaves schools with two choices: frantic coverage or thoughtful focus. Both choices are fraught with risk, with the former risking learning and the latter risking omission of some standards. The evidence supports those who take the second risk.

The entire standards movement is put at risk by those who have replaced focus and meaning with quantity and coverage. One must not, however, confuse effectiveness with popularity. When one suggests that the memorization of the dates of the Crimean War may be less meaningful than a deep understanding of national economics and international conflict, one is besieged by those who equate intimacy with the Crimean War with cultural literacy. When one suggests that students might well focus on measurement, calculation, and problem solving rather than the lower level distinctions between the trapezoid and the rhombus, one is reviled by the partisans of the rhombus. Even the stegosaurus has its lobbyists who are armed for battle when one suggests that students are better served by immersion in the scientific method than by one more diorama in which prehistoric creatures are displayed with craftsmanship that outstrips paleontology.

The solution to the triumph of superficiality is the concept of power standards, that small subset of standards that must pass three crucial thresholds. First, do the knowledge and skills implied by this standard endure? That is, will they last beyond a few test questions in a single year? The scientific method endures, whereas the stegosaurus, alas, does not. Second, does the standard have leverage? That is, is it applicable across a wide spectrum of other standards? For example, the facility of a student to create and draw inferences from tables, charts, and graphs has extraordinary leverage, as these skills are found not only in the mathematics standards but also in science, social studies, and language arts. Similarly, the ability to write a descriptive, expository, and persuasive essay crosses the boundaries of many disciplines. Third, is the standard required for the next level of instruction? This threshold would require a focus in middle school on calculation, problem solving, measurement, and a few other areas. Most high school algebra teachers would be delighted if students came to them proficient in these few power standards, even if their young charges are rhombus impaired.

The best way to identify power standards is to ask faculty members this question: If you were giving advice to a teacher in the next lower grade, what knowledge and skills would that teacher need to impart to students in order that they enter your class next year with confidence and success? This question has been asked of hundreds of teachers at all levels of instruction, and few have replied that the teacher in the next lower grade must cover the standards. Rather, educators universally provide responses with two characteristics: brevity and balance. The lists are invariably short, rarely comprising more than 10 items. Moreover, the lists are balanced, including not only discipline-specific information but also a healthy dose of literacy, organization, time management, and appropriate behavior. It is the wisdom of these teachers, not the mindless pursuit of coverage, that leads to success in future classes and on high-stakes tests. Another question may help to disarm those who remain advocates of coverage because the "rhombus might be on the test": When you took your last high-stakes test—any in the alphabet soup of tests in your past including the GRE, ACT, SAT, and PSAT—did you prepare for the exam by attempting to memorize the item base of the Educational Testing Service? Serious reflection suggests that a focus on a few skills was the key to success at the time when the coverage advocates were students themselves. Power standards, not frantic coverage, form the answer to both successful educational practice and improved test performance.

Transformation 4. From Seniority to Equity in Teacher Assignment
Without question, the most controversial change suggested here is a challenge to the tradition in which the seniority system is used to allocate the greatest teacher experience, certification, and quality to schools with the wealthiest students. The data are unambiguous on this point; Haycock (1998) referred to a synthesis of the evidence when she noted that the impact of subject-matter certification is far more significant than demographic variables on student performance, yet around the nation the allocation of teachers with subject-matter certification is markedly inequitable. Poor and minority students are far less likely to have certified teachers than their economically advantaged peers. Within districts or even individual school buildings, the inequities are pervasive:

teachers with greater experience, degrees, and certifications are assigned to students in advanced classes, whereas the poorest and least able students frequently find themselves in classes led by long-term substitute teachers or those new to the profession. To be sure, some new or uncertified teachers are extraordinarily competent, whereas some veteran, certified teachers are extraordinarily incompetent. Further, these generalizations should not diminish the heroic efforts of the many teachers with multiple certifications, advanced degrees, and years of experience who labor in the most disadvantaged and challenging schools. On the whole, however, the certification, degrees, and experience of teachers are highly associated with student success, and teachers with those characteristics in abundance are overwhelmingly assigned to students who are economically advantaged (Rothstein 2000).

There are some notable exceptions. Some schools in California, Texas, and Illinois take pride in their policy of equitable assignment of teachers. Within those schools, every teacher has the same opportunity to work with the most and least able of the students. These schools, however, represent the exceptions. The typical reaction to the suggestion that school systems replace seniority with equity is the question, "Do you mean that a 30-year veteran with a master's degree should have to teach basic math while a new teacher should teach AP Calculus?" The answer is, as a matter of simple justice, yes. The argument is not that all veteran teachers be reassigned to poor schools, but only that educators provide equal educational opportunity for all children. This should not be an excessively demanding standard for a nation and a profession claiming a commitment to equity.

The predominant litigation in school finance over the past few decades has focused on the equitable allocation of financial resources. Although this is important, financial resources are not sufficient. The most important resource any educational system has to allocate is the expertise of its teaching professionals. The assignment of teachers to different schools and courses, decidedly more sensitive and less convenient than the reallocation of financial resources, has the greatest opportunity for a profound impact on student equity.

Transformation 5. From High Stakes to Meaningful Evidence
By 2003, 26 states will use some version of a high-stakes test as a criterion for a high school diploma (Coleman 2000). Despite the apparent popularity of high-stakes tests with educational policymakers, there remain two significant problems: students may be proficient in state standards and fail these tests, and students may succeed on the tests without being proficient. Indeed, the two errors feed on one another, as fear of the first error (and the attendant threat of litigation) leads to a greater likelihood of the second error.

The quest for the holy grail of testing—the perfectly (or even reasonably) reliable and accurate large-scale test of student competence that in a few hours can sum up 13 years of public schooling—is destined to fail. This pessimism is not based entirely on the sorry record of large-scale standardized tests (Lemann 1999; Popham 2000), though the history is hardly distinguished. Although validity, reliability, alignment with the curriculum, and

the opportunity for students to succeed are the criteria that would make a high-stakes test defensible, these criteria cannot be present simultaneously.

For a test to be valid, it must test what it is intended to test. That means, for example, that tests of writing test the ability of the student to write, rather than the student's knowledge of a specific content area suggested by a writing prompt. The larger the body of writing examined, the more likely that one obtains a valid estimate of the ability of that student to express a coherent thought in writing. The smaller the body of writing examined (in the extreme case, when the ability of a student to write is estimated on the basis of on a single writing sample), the more likely that the inference drawn about that student is representative of the content rather than the written expression. In brief, smaller samples of writing are more likely to yield invalid results. In the context of mathematics, science, or social studies, the requirement for validity conveys the notion that assessments are testing what the student has been led to believe is a proficient understanding of these disciplines. Many state standards, however, require deep understanding rather than mere factual recall. Deep understanding, whether of the scientific method or of the analysis of historical trends or of the challenges of Internet-based research, is not revealed in isolated tests. The very words of the standards require evidence that extends far beyond the score on a single test.

For a test to be reliable, it must evaluate student performance consistently. In the case of multiple-choice tests, this implies a consistency among items. If a student answers 9 out of 10 arithmetic problems correctly and misses the tenth problem, then either the last problem was quite difficult or, if other similarly able students got the tenth problem correct, then that test item was unreliable. It is as if the same student mounted a scale 10 times, and for the first 9 readings the student's weight was pronounced healthy and on the tenth reading the student's weight was labeled unhealthy. Because the same characteristic was being measured and the student had presumably not changed much between the ninth and tenth reading of the scale, it is the scale that is the problem, not the student.

The essence of the problem is that the requirement for validity implies multiple tests with some variations in administration and with independent judgments rendered regarding the ability of the student. The requirement for reliability requires a single instrument, consistently administered. Think about it. Most standardized tests fulfill only the latter requirement: single instruments, consistently administered. That is why college admissions tests such as the SAT and ACT are considered reliable. Few observers, however, continue to cling to the fiction that those tests are valid—that is, that they accurately predict how students will perform in college (Rooney and Schaeffer 1998).

A solution to the validity-reliability conundrum relies on the abandonment of the illusion that a single testing instrument such as the individual high-stakes test is the appropriate or accurate (or reliable or valid) measurement of student achievement. Student mastery is best evaluated when it is based on a body of evidence collected throughout an academic career, rather than on a single score. The consequence of the failure of a student to provide evidence of success should not be a lifetime of discredit and an eternity of litigation with the school system. Rather, the consequence of insufficient evidence of

student proficiency should be another opportunity to provide that evidence. Indeed, it is the feedback on the first attempt that informs the second attempt. It is the transparency, not the secrecy, of an assessment that makes it useful to students and believable by the public.

The use of a body of evidence rather than a single assessment score to ascertain student mastery is certainly more cumbersome and labor-intensive. Despite the logistical challenges, many school systems are succeeding. High schools from the Rocky Mountain School of Expeditionary Learning in Denver, Colo., to schools in Middletown, R.I., and North Clackamas, Ore., to more than a dozen schools in Wisconsin all require evidence that is based on multiple assessments with multiple opportunities for student success as part of their high school graduation criteria. This is the only meaningful way to evaluate students, particularly when the consequence of inadequate assessment is the denial of a high school diploma.

The use of a body of evidence to analyze student achievement does not negate all uses of standardized tests. Although it is politically popular to excoriate these tests, it is more useful for school administrators to understand the limits and strengths of such assessments. The critical issue is not the tests themselves, but how the information from those tests is used. On the one hand, a sound assessment used unwisely to make inappropriate decisions leads to the classic case of a valid and reliable test used to make invalid and unreliable policy. On the other hand, tests can provide vital information for school leaders and policymakers, and the information yielded from those tests should not be casually dismissed merely because the results are unpopular or embarrassing.

The Endurance of Standards and Civil Discourse

Will standards become another passing fad in American education, or will the standards movement endure over time? If standards are to endure, then educators must get beyond test prep and embrace thinking, reasoning, and communication. They must make grades accurate and meaningful. They must replace frantic coverage with power standards. They must commit to the principle of equity in teacher assignments. Finally, they must move from single high-stakes tests to a body of evidence as the measurement of the degree to which students meet standards.

The transition of academic standards from passing fad to permanent educational fixture will be achieved with enormous difficulty. The controversy surrounding standards will continue for some time, but surely both defenders and critics of standards can set an example for students by elevating the discourse and conducting the debate with greater civility and respect.

References

Coleman, A. L. 2000. None of the above. Education Week 19 (43): 42, 45.

Groves, M. 2000. California schools show big gains in test scores. Los Angeles Times, 5 October.

Kohn, A. 1999. The schools our children deserve: Moving beyond traditional classrooms and "tougher standards." Boston: Houghton Mifflin.

Haycock, K. 1998. Good teaching matters: How well-qualified teachers can close the gap. Thinking K–16 3 (Summer): 1–16.

Lemann, N. 1999. The big test: The secret history of the American meritocracy. New York: Farrar, Straus & Giroux.

Marzano, R. J., J. S. Kendall, and L. F. Cicchinelli. 1998. What Americans believe students should know: A survey of U.S. adults. Aurora, Colo.: Mid-continent Regional Education Laboratory.

McNeil, L. M. 2000. Creating new inequalities: Contradictions of reform. Phi Delta Kappan 81 (10):728–34.

Ohanian, S. 1999. One size fits few: The folly of educational standards. New York: Heinemann.

Popham, W. J. 2000. Testing! Testing!: What every parent should know about school tests. Boston: Allyn and Bacon.

Reeves, D. B. 1997. Making standards work: How to implement standards-based assessments in the classroom, school, and district. Denver, Colo.: Advanced Learning Press.

————. 2000. Accountability in action: A blueprint for learning organizations. Denver, Colo.: Advanced Learning Press.

Rooney, C., and B. Schaeffer. 1998. Test scores do not equal merit: Enhancing equity and excellence in college admissions by de-emphasizing SAT and SAT results. Cambridge, Mass.: FairTest.
Rothstein, R. 2000. Getting good teachers for poor schools. The New York Times, 20 September, national edition, sec. A, p. 21.

St. Louis Post-Dispatch. 2000. Editorial, 14 September.

Wiggins, G. 1988. Educative assessment. San Francisco: Jossey-Bass.

State accountability requirements,

teacher independence,
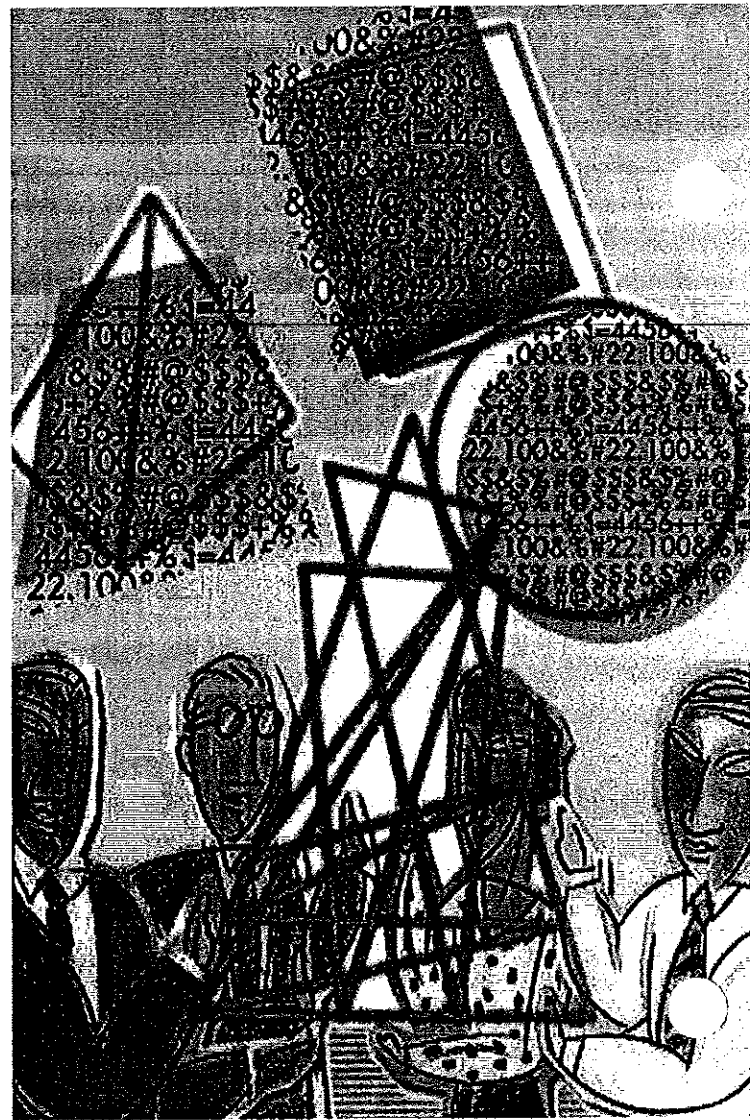
and local school board policy

Caught in the
MIDDLE

By Douglas B. Reeves

School board members are caught in the middle of one of the most intense political and personal debates in recent decades: the debate over standards and assessment. The argument is characterized by extremes, with state policymakers uniformly demanding high standards at the same time that increasingly militant teachers are criticizing high-stakes assessments and demanding more autonomy in the classroom. The two sides are so far apart on the fundamental issues that the most well-intentioned school board member will have a difficult time reconciling them. But board members can take heart: There is a way to reconcile state demands for accountability with teachers' demands for a reasonable degree of independence.

All 50 state legislatures hold a significant amount of authority over the purse strings for public education, and more than 40 states tie funding to state-imposed accountability systems. The states are making multibillion dollar investments and, they reason, it is only fair and reasonable to expect accountability for those expenditures. Legislators believe it is only reasonable to ask, "To what extent are we meeting the academic standards that we have established over the past decade?" Reasonableness, however, gives way to an extreme view when "accountability" is interpreted as nothing more than a single test score. It's as if a physician were to decide whether to operate on a patient on the basis of the patient's body temperature. That reading might be one important indicator of human health, but no one would characterize body temperature alone as a sufficient basis on which to make life-or-death medical decisions. Such a conclusion would be superficial and extreme.

One person's extremism is another's reasonableness, however. Consider the arguments of those opposed to standards, assessment, and accountability. They go beyond challenging the excessive breadth of many state standards and the flaws in many state tests, both of which are reasonable arguments. "Leave teachers alone and let them teach!" That is the rallying cry at an increasing number of regional and national teacher and administrator meetings. The resistance to higher academic standards and demanding assessments is well-organized, militant, and visible. In fact, it is possible that your school district is paying to send teachers and administrators to conferences where participants are urged to refuse to prepare students for state tests and to encourage students to sabotage the tests.

In the middle of this fractious debate are school board members, who must respond to the state mandates while considering the wishes of the school staff. Teachers unions are powerful entities in most school systems; more to the point, the unions and professional associations offer some valid and pointed objections to the manner in which educational accountability has been used as the blunt instrument of educational reform by most states. And increasingly, their arguments are being taken up by parents and other members of the public.

In a debate of extremes, where is the middle ground?

Accountability—physical or autopsy?

To reconcile the competing issues in this debate, school board members must first determine the purpose of educational accountability. Is it a physical or an autopsy? Most state-level ac-

countability systems gather only test scores, with a passing reference to other variables—typically poverty and ethnicity. This leads casual observers to rely on a single instrument—high-stakes tests—to evaluate schools and teachers. Observers make a mistake when they do this. After all, if the subject were our health, an autopsy would be of little help. We don't want to know why we died; we want to know how to get better.

If educational accountability is to become a physical rather than an autopsy, what information do we need? Clearly, test scores can be part of this equation, but not the sum of the matter. In research conducted at the International Center for Educational Accountability, we have focused on what we call the antecedents of excellence—those variables at the school and classroom level that are associated with improved student achievement. The results of our research suggest that accountability can be comprehensive, constructive, and meaningful, thereby bridging the gap between state accountability systems and teacher autonomy.

For educational accountability to be comprehensive and meaningful, it must differ in fundamental ways from the typical litany of test scores—the "effect variables" of the educational system. A comprehensive and meaningful accountability system, by contrast, will focus on "cause variables"—those specific strategies at the classroom and school level that are associated with improved student achievement. For example, a typical accountability system might detect an increase in test scores and speculate that the increase was associated with some other variable that happened to be measured. Typically, this results in a focus on demographic characteristics of students. More poverty, the superficial analysis goes, equals lower student achievement. Higher percentages of minority student enrollment, by the same simplistic analysis, are also inexorably associated with lower achievement.

The 'two-variable fallacy'

The problem with most analyses of educational accountability stems from a classic error in logic: the assumption that, when two events happen at the same time, one caused the other. This so-called two-variable fallacy leads to the mistaken conclusion that poverty *causes* lower test scores or that minority enroll-

---

For further reading

Ladd, H.F. *Holding Schools Accountable: Performance-Based Reform in Education.* Washington, D.C.: Brookings Institution, 1996.

Mathers, J.K. *Educational Accountability Systems in 50 States.* Denver: Education Commission of the States, 1999.

Reeves, D.B. *Accountability in Action: A Blueprint for Learning Organizations.* Denver: Advanced Learning Press, 2000.

Schmoker, M. *Results: The Key to Continuous School Improvement.* Alexandria, Va.: Association for Supervision and Curriculum Development, 1999.

---

ment *causes* lower test scores. Fortunately, these errors are easily exposed once a school system is willing to measure other variables. In fact, in both local and national studies, it is clear that poor schools with high percentages of minority enrollment are far less likely to have teachers who are fully certified in their subject matter. The difference, it turns out, is not so much in the students as it is in the teachers. Long-standing agreements and traditions have allowed teachers with more experience and credentials to select the school in which they will teach. These arrangements have led, inevitably, to a disproportionate number of our most-qualified teachers teaching our least-needy students, and our least-qualified teachers teaching our most-needy students. But because of the two-variable fallacy, observers don't consider teacher qualification and certification, assuming instead a relation between only two variables—demographics and achievement. At worst, this is institutional racism; at best, it is sloppy research indeed.

A growing body of research—including work by the Education Trust in Washington, D.C.; William Sanders at the University of Tennessee; and my colleagues at the Center for Performance Assessment, with which the International Center for Educational Accountability is affiliated—makes it clear that poverty and ethnicity are not the primary causal variables related to student achievement. These demographic variables have strength only when researchers fail to measure teaching and leadership variables. In other words, when the adults in the system—teachers and leaders—start to take responsibility for their role in educational accountability, it becomes much more difficult to blame children and parents for poor student achievement.

Reconciling the differences

How can school boards reconcile the differences between the test-dominated accountability systems and the growing backlash against standards and assessments from teachers and parents? How can school boards avoid the "two-variable fallacy" without imposing enormous bureaucratic burdens on the schools? The answer lies in a comprehensive accountability system. Such a system does not ignore state demands for improved achievement as measured by test scores, but it does insist that state test scores must be placed in context. Comprehensive accountability systems gather information not only on test scores, but also on the antecedents of excellence—the specific professional practices at the classroom and school level that might or might not be associated with improved student learning.

Examples of these practices that can be measured with local comprehensive accountability systems include the frequency of writing assessments and the use of traditionally nonacademic subjects, such as physical education, music, and art, to improve student performance in essential academic skills. This can be displayed on a bar chart that shows the frequency of such interdisciplinary assessments. In addition, a comprehensive accountability system can measure substantive parental involvement, technology applications in the curriculum, and the

relationship of community service and extracurricular activities to improved academic achievement.

The results can be surprising. Some of the most popular academic programs might have minimal impact on student achievement, while some overlooked and pedestrian initiatives might be associated with significant improvement in results. For example, in Riverview Garden, Mo., the district tracked the frequency of writing assessments in every class, including physical education, art, and music. The percentage of students who were proficient or better in writing tripled during the 1999-2000 school year. Moreover, the district experienced striking gains on state test scores. The relationship between the improvement in classroom-level writing assessments and improved state test scores would not have been apparent unless the district tracked both test scores and classroom practices. Comprehensive accountability, in other words, is the antidote for the "fact-free debate" that so frequently characterizes deliberations among school leaders and policymakers.

Comprehensive accountability in practice

Comprehensive accountability is not the pipe dream of an academic. It is used in hundreds of schools in Wisconsin, Indiana, Missouri, Colorado, and California, schools we have worked with at the International Center for Educational Accountability. While each system reflects the unique needs of the community, all of them have in common two critical characteristics. First, each includes both districtwide and school-based accountability indicators. The districtwide indicators reflect the state requirements for an emphasis on test scores and typically also include measurements of attendance and dropout rates. School-based indicators, by contrast, are the "cause variables"—measurements of specific instructional, assessment, and leadership practices that can vary widely from one school to another. By liberating each school to select different school-based variables, the school board has honored the demand for independence and creative thought at the school level. The key to the system is that every school-based indicator must provide support for the systemwide indicators.

The second critical characteristic of successful comprehensive accountability indicators is the use of qualitative information. As much as I love statistics, I must confess that not everything can be measured with a number. In fact, quantitative data and qualitative information are not in opposition to each other. In the realm of educational accountability, the qualitative information—typically a narrative of school climate, achievements, triumphs, and tragedies—creates the lens through which we can better understand quantitative information. These narrative descriptions are not extensive. They usually take the form of a one-page description that allows the casual reader or board member to better understand the context of test score data.

The school board and accountability

The school board need not spend an extraordinary amount of time developing and reporting accountability indicators. The board should, however, perform two critical functions. First, in conjunction with the superintendent, the board should create an accountability task force that is broadly representative of the community. Second, the board should use the task force not only as a body that reports test score data, but also as an analytical body that can provide independent analysis and information to the board and to the public. Anyone can get test scores from the newspaper. What the board rarely receives is additional insight into which specific programs and strategies work and which ones do not work.

The task force can systematically analyze information and give the school board insight into the most effective—and least effective—practices in the district. It might be the first time that many school systems have had the tools to determine, beyond a vague feeling, what strategies are most effective in a local environment. That can be invaluable in the board's decision making. Sales pitches and advertisements for one program or another are always littered with the phrases "research-based" and "proven results." These claims are usually based on national evidence that might not be applicable to a particular locale. Policymakers and school leaders need to have these assertions confirmed with local evidence. Specifically, they need to be able to systematically test the hypothesis that a program or technique that has worked elsewhere also can be effective in local schools. An effective accountability task force can conduct this sort of inquiry and provide independent information that is constructive, accurate, and relevant to local needs.

The heart of a comprehensive and effective educational accountability system is its impact on student achievement. Without better accountability information, board members will know only that test scores are not adequate but will have little insight into how to improve them. Contrary to the prevailing accountability mythology, a good accountability system is not distinguished by its ability to award simplistic grades to schools based on test scores. Effective accountability systems must give board members and school leaders vital information not only about school performance, but also about the antecedents of excellence—what strategies work and what strategies do not.

In most school districts and states, the accountability discussion centers on test scores and demographics. The resulting blame and aggravation do not, to put it mildly, advance the cause of reason in educational debates. The better alternative of comprehensive accountability maintains the public's legitimate concern for effective performance reporting while providing educational leaders and policymakers the information that they need to make wise decisions about the future of our schools.

Douglas B. Reeves (dreeves@testdoctor.com) is the author of *Accountability in Action: A Blueprint for Learning Organizations* and the president of the Denver-based International Center for Educational Accountability, http://www.edaccountability.org or (800) 844-6599.
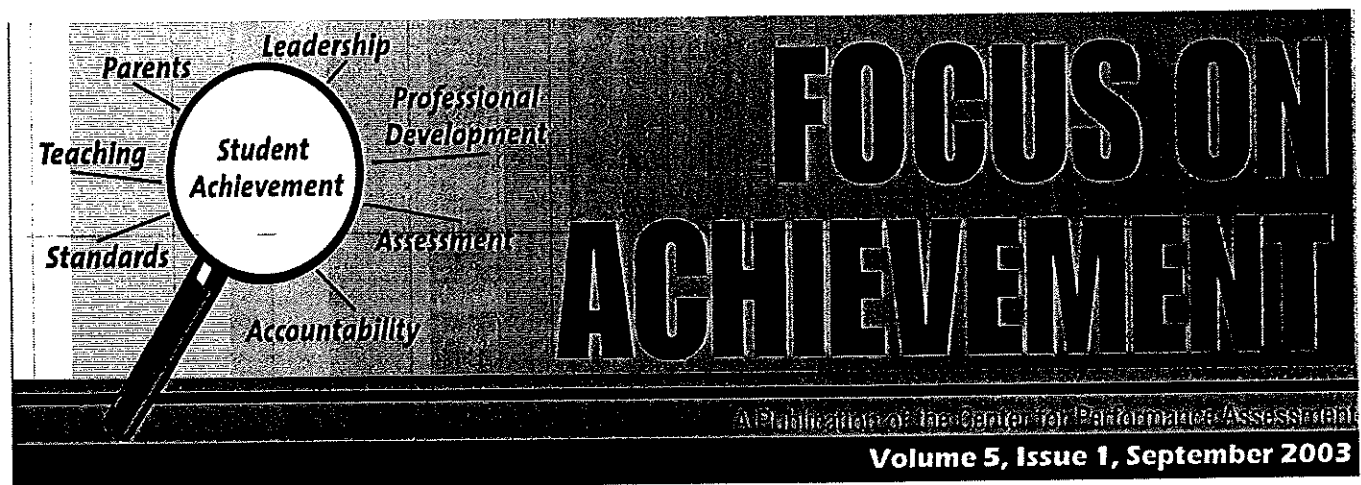
Data Without Tears — Accountability As A Treasure Hunt

By Douglas B. Reeves

"You can't be serious," the principal scoffed. "You are asking us to look at the data for each individual teacher? I'll have a revolt on my hands — we're just not ready for that. Besides, with budget cuts and increasing test pressure, anything I try to do on the subject of educational accountability will just push my teachers right over the edge."

This school leader is a decent and committed professional who loves and cares about kids. But because of a disappointing history using test data, the resistance to any constructive form of accountability is overwhelming. Chances are, the teachers in this school also have a negative history in using test data, with anger and defensiveness the typical result of an examination of test scores. After all the excuses have been given, the principal might as well have said, "I'm all for data-driven decision-making — I just don't want to look at the data."

Find the Source of Success

How can we stop this destructive cycle in which the use of test data is associated with anger and recrimination? First, principals and teachers must reframe the conversation around student data as a "treasure hunt." The first question any principal in any school must ask is, "Why are we so good in the following areas...?"

This may not be an obvious question in those schools where test scores are stagnant and the averages are low. Nevertheless, effective leaders peel away each layer of data, finding the sources of success. Even when reading scores are disappointing, an individual teacher may have a gain in a subscale, such as vocabulary. Even when math scores are low, a teacher may have been successful in a subscale, such as measurement.

Rather than bemoan the low reading and math scores, effective leaders ask, "Why were we successful in this particular subscale?" Effective analysis of data is a treasure hunt in which leaders and teachers find those professional practices — frequently unrecognized and buried amidst the test data — that can hold the keys to improved performance in the future.

When the process of data analysis is converted into a treasure hunt, an air of gloom and defensiveness is replaced with an environment of excitement, anticipation, and pride. We know that effective feedback for students requires that the feedback is accurate, specific, and timely. The same is true when it comes to effective feedback for adults. The most effective schools do not engage in data analysis as an annual review of test scores, but as a continuing endeavor. It is simply the way they do business, referring to external and internal data, finding their most effective instructional practices, and using data to provide accurate and effective feedback for teachers.

Replace Gloom with Pride

The most important consideration in data analysis is this: Teachers and school leaders want to be successful. We have nothing to fear from educational accountability that is accurate and constructive. We have everything to gain from data analysis that will help us to recognize and replicate our own best practices.

In future issues of *Focus on Achievement,* we will consider what indicators can be measured at the school level and how accountability can be used to improve professional development.

The Center for Performance Assessment provides a number of resources, including free web-based downloads, to help you in your accountability and data analysis efforts. You can also receive direct and personal assistance by going to the web site or calling our toll-free number. We must all live with accountability — the only issue is whether we will settle for a list of test scores or whether we will, using constructive and holistic techniques, get accountability right.
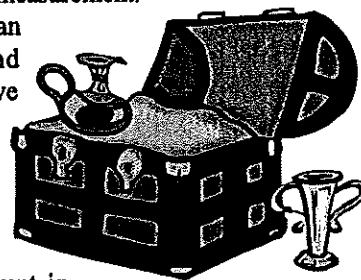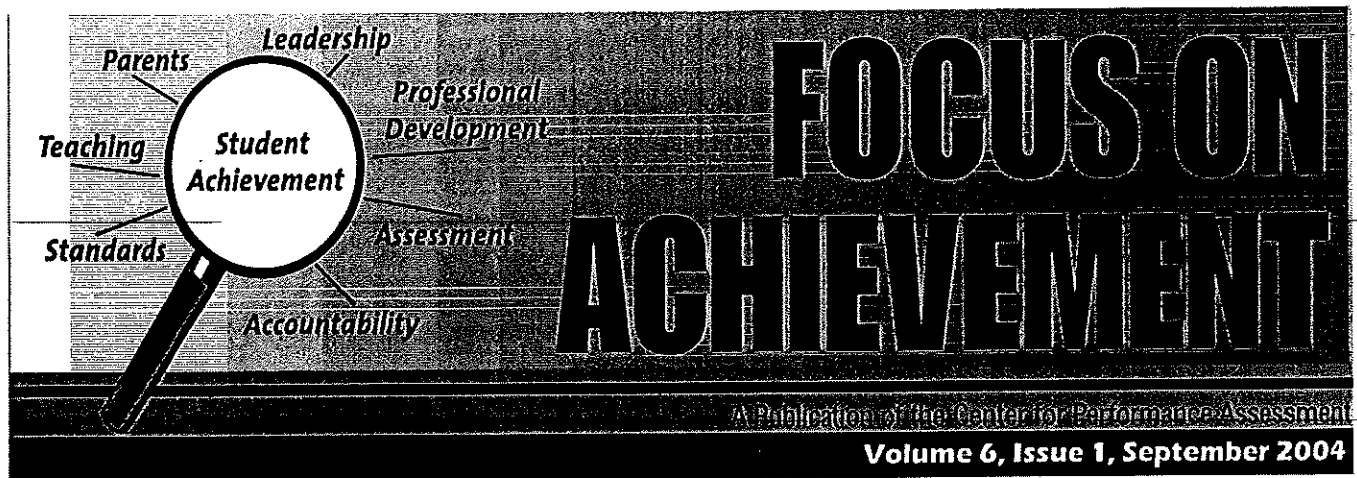
Accountability for Learning

Keys for Transformational Schools

By Douglas B. Reeves, Ph.D.

The theme for the 2004-2005 issues of *Focus on Achievement* will be Accountability for Learning: Keys for Transformational Schools. In this series of articles, we will explore ways to make accountability a constructive force that can transform schools and entire districts.

The good news is that we know what works. Districts throughout the world have demonstrated that high-performing schools can become even better and low-performing schools can make dramatic improvements. The issue now is execution – implementing the leadership and teaching steps necessary for student success.

Breaking Through the Achievement Plateau

As scores for the previous school year are received, educators and school leaders are caught in a bind. If the scores are low, there is a pervasive sense of gloom. "Our staff is at a breaking point and they have already worked as hard as they can, stretching every dollar, hour, and person to the limit." If scores are high, on the other hand, there is a barely perceptible celebration, as any congratulatory language is quickly swept aside by the question, "What have you done for me lately?"

The most common phenomenon, however, is that scores are flat. During the first few years after the implementation of standards and accountability, curriculum, collaboration, and interventions were implemented with determination. Student achievement improved, even in the most challenging schools.

But recently, the air seems to seeping out of the balloon. The pervasive feeling is that "We did that *last* year – can't we finally relax?" If we are to break through the achievement plateau, then our strategy must be more refined than exhortations of "higher scores!"

Beyond "Higher Scores"

While the measurement of student achievement on state test scores is undeniably important, the measurement of test scores alone is insufficient to guide teaching and learning. Test scores represent the effects of teaching, curriculum, leadership, parent involvement, student engagement, and other important variables.

If we only measure the effects without also considering the causes, it is as if we were to attempt to deal with the health concerns of an overweight student only by looking at the scale. Even if the student lost weight, we would not know if the causes were appropriate diet and exercise or anorexia and drug use. The "score" does not reveal sufficient information to improve the health of the student.

Keys to Transformation

The Center's work in all fifty states and on five continents suggests three keys to successful transformation. First, accountability systems must identify and document best practices. This is not the same as popular practice, but rather those specific strategies in teaching, curriculum, and leadership that are specifically linked to improved achievement and equity. An accountability system that is focused only on scores but not on professional practices will fail to achieve this essential first step.

If I could identify a single best practice for schools, it would be the "science fair for adults" in which every school lists not only its test scores, but the specific strategies in teaching, curriculum, anu

> **If I could identify a single best practice for schools, it would be the "science fair for adults" in which every school lists not only its test scores, but the specific strategies in teaching, curriculum, and leadership that were associated with those scores.**

Accountability in Action, 2nd Ed., Released with Updated 90/90/90 Schools Research Data

Accountability in Action is a practical, step-by-step guide to the creation of a comprehensive accountability system. This book provides a bridge between standards implementation and the system that will help schools assess the effectiveness of their efforts.

Dr. Douglas B. Reeves demonstrates how policy makers, leaders, teachers, parents, and students can create a comprehensive accountability system to implement teaching and learning programs that help *all* students succeed.

Each chapter includes discussion questions and ready-to-use worksheets to guide educational teams through the process. Appendices include sample reports, nameplates for system design, checklists, indicators, and resources.

***This book helps your school and district:***
- Assess current accountability practices
- Evaluate student achievement using multiple measures
- Identify program strengths and weaknesses
- Design accountability measures
- Implement comprehensive accountability systems
- Improve communication with educational stakeholders

What major topics are addressed?
- Accountability: the key to sustained reform
- Accountability system design
- The role of standards in educational accountability
- Leadership and policy in educational accountability

Who should use this book?
- School board members
- Educational leaders
- Research, assessment, and accountability specialists
- Classroom teachers
- Staff development professionals
- Parents and community and business leaders
- Higher education faculty
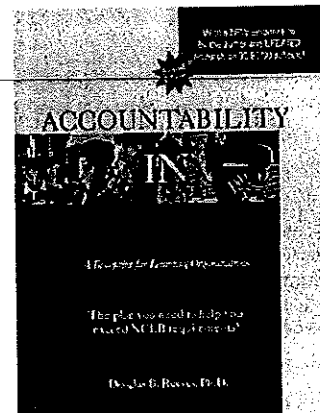
ISBN 0-9747343-1-4

Keys to Transformation
Continued from Page 1

leadership that were associated with those scores.

Second, we must bring our best practices to scale. In almost every school the Center has served, we find teachers and school leaders doing outstanding work. Unfortunately, this outstanding work is provided to a fraction of students who need it.

There are wonderful reading intervention programs provided for 15 percent of students, but 40 percent of students need it. There are outstanding math programs provided for 30 percent of the students, but 80 percent need it. There are enrichment and advanced learning opportunities provided for 5 percent of the student population, but 50 percent or more could benefit from the same opportunities if only the district brought their best practices to scale.

These teachers and leaders know what do to. Their challenge is to recognize their own best practice and bring it to scale.

Third, we must conduct a "garden party." The metaphor of the garden party is taken from the book *Making Standards Work* in which I suggested that we must "pull the weeds before we plant the flowers." Schools are taking this encouragement seriously, dramatically reducing or eliminating such traditional practices as announcements (particularly those that are ignored by everyone except the person holding the microphone), assemblies (particularly those that fail to improve the community-building or achievement objectives of the school), and faculty meetings (particularly those that do no more than deliver announcements).

In every school and classroom there are weeds to be pulled – at least a single meeting, activity, unit, chapter, subchapter, and (dare I say it?) holiday. If you are not willing to pull the weeds, you will never plant the garden. If you are not willing to save teachers and school leaders the hours invested in unproductive activities, you will not bring your most effective practices to scale.

In future issues of *Focus on Achievement* we will provide examples of schools and systems that are implementing these principles of success. If you have a best practice to share, please send it to us at info@makingstandardswork.com so we can give your best practices the attention that they deserve.

> **More Information about Accountability for Learning and Transformational Schools**
>
> *Accountability for Learning: How Teachers and School Leaders Can Take Charge*, Douglas B. Reeves, ASCD, 2004
>
> *Accountability for Greater Student Learning*, featuring Douglas B. Reeves, Video Journal of Education, 2000
>
> *Accountability in Action: A Blueprint for Learning Organizations*, Douglas B. Reeves, Advanced Learning Press, 2000-2004
>
> *What Works in Schools: Translating Research into Action*, Robert J. Marzano, ASCD, 2003

Holistic Accountability: Serving Students, Schools, and Community

Reviewed by Cecilia M. Di Bella, Ed.D.

**HOLISTIC ACCOUNTABILITY:
SERVING STUDENTS, SCHOOLS,
AND COMMUNITY**
By Douglas B. Reeves
c. 2001, Corwin Press,
Thousand Oaks, CA
ISBN: 0761978321

Accountability: The A-Word for Schools in the Twenty-first Century

School administrators, teachers, and school board members must come to grips with the simple fact that the demand for greater educational accountability is here for the long haul. In truth, parents, legislators, taxpayers, and students have every right to expect, and demand, that funds allocated for education are expended prudently, appropriately, and effectively. The tipping point in this mandate is who defines the expected results and how the results are defined.

Presumably, the purpose of any educational accountability system is to improve student achievement. However, this simple premise belies the exclusive use of test scores or quantitative data as the primary measure of a student's progress or accomplishments. Leaving aside Howard Gardner's theory of multiple intelligences, only someone with a doctrinaire approach to student achievement would overlook the accomplishments of students whose artistic, technological, athletic, or musical abilities exceed their standardized state test scores.

If colleges and universities across the country are moving away from reliance on SAT scores to a wider array of factors for admission, why are public schools so relentless in their demands for such a narrow definition of student achievement? Or as public purse strings

tighten, are simplistic, quantifiable data an easy and convenient response to the question, "Are students learning?"

This question leaves little room to embrace the complex milieu in which schools and student learning occur or discussion of the more appropriate issue, What are students learning? And herein lies the battle line: those who rely on test scores as the sole measure of accountability (i.e., did federal or state funds result in higher scores) versus those who subscribe to a broader view of accountability.

Enter Douglas Reeves, who views accountability as an "evolving arena fraught with dangers but rife with opportunities for school leaders willing to approach it in a constructive manner." He acknowledges the importance of test scores in accountability systems but emphasizes that scores must be considered in context. Reeves stresses the significance of the context in which instruction occurs, namely the instructional practices, standards, curriculum, leadership techniques, and resource allocation. In other words, accountability must be holistic rather than fragmented.

In *Holistic Accountability*, Reeves succeeds in recasting the accountability dilemma into a win-win plan for school administrators who are weary of the test score sweepstakes and for policy makers who are demanding results. In essence, he redefines the parameters of accountability systems, moving the concept closer to a strategic planning process with both quantitative and qualitative goals and corresponding measures of achievement.

The four pillars of his approach are structure, collaboration, implementation, and communication. And it is an excellent approach—one that is sound,

reasonable, comprehensive, and relevant. More important, its components satisfy and comfort all parties.

The book includes case studies of three school districts that have followed a slightly modified version of Reeves' approach: Wayne Township, Indiana; Milwaukee, Wisconsin; and Riverview Gardens, Missouri. The accompanying sample accountability reports are very useful, for they illustrate how an administrator can steer accountability away from the narrow view to a more comprehensive wide-angle view of the district's program and student development.

The reports are a thoughtful and meaningful look into the improvement plans and goals for many areas of the school system, not simply test scores. In addition, Reeves' approach involves the whole community, not simply the school staff—a technique that is bound to build community-wide consensus and support.

Policy makers and school administrators who are frustrated with the scorecard version of accountability will find *Holistic Accountability* refreshing and satisfying. Happily, it is free of jargon and succeeds in demystifying and recasting one of the most contentious areas in public education today.

School business administrators will also find the book an excellent primer about accountability and may draw a parallel with their own efforts to move beyond viewing fiscal data not as a documentation of what was spent, but how expenditures improved student achievement. ■

Cecilia M. Di Bella, Ed.D., is superintendent of Nahant School District in Nahant, MA. She is also a member of ASBO International's® Editorial Board.

school administrator

search site

awards and scholarships

career center

conferences

education marketplace

government relations

issues and insights

links

membership

publications

state associations

home

Holding Principals Accountable

Seven considerations for effectively evaluating your site administrators
BY DOUGLAS B. REEVES

The clarion call for accountability falls disproportionately on superintendents and principals. Boards of education, frustrated that student achievement has not sufficiently improved, are elected based on promises of change, and leadership positions are the most visible indicators of apparent stagnation.

Thus the promise that "achievement will improve or heads will roll" sums up a remarkably facile approach to the complex challenge of educational leadership. Only an effective leadership accountability system can respond to this counterproductive combination of threat and bluster.

Accountability is more than a promise of change. Properly implemented, it provides a clear strategic direction for schools and their leaders, with a laser-like focus on student achievement and specific strategies that will be used at the school level to create improvement. If the promise of accountability is to be achieved, however, then more clarity and less rhetoric must be the order of the day.

Counting Improvement
After reviewing leadership evaluations and accountability systems in hundreds of schools, I have concluded there are seven keys to effective educational accountability systems: balance between achievement and improvement, specificity, focus on student performance, frequency, adaptation to individual strengths, rewards for the tough choices and reflection.

Even the most perfect accountability system, however, will fail if leaders are not given authority commensurate with their responsibility.

feature
department
editor's note

advertising
author guidelines
contact us
editorial calendar
faq
past issues
photo examples
purchase copies
reprints &
permissions
subject index
magazine home

E-mail this article

77

In many districts, site administrators have little or no control over the hiring and discipline of teachers, the daily instructional and assessment activities within the classroom and the availability of basic learning resources, including textbooks, computers and even desks. Accountability is a sham when superintendents and principals are subject to public humiliation and career jeopardy when they fail to improve student achievement under these circumstances.

If--and only if--you're ready to give building leaders the ability to carry out their mission, here are some effective ways to hold them accountable.

- *Balance achievement with improvement.* The most consistent complaint I have heard from school administrators around the country is that they are being held accountable for things over which they have no control. One principal in Wisconsin spoke for thousands when he said, "I don't feed them before they get here, discipline them at home, limit their consumption of alcohol and drugs except for six hours a day or give them a home where they are safe from violence and abuse. But every one of those factors affects their performance at school, and you're going to tell me that it's all my fault?"

  It doesn't take most principals long to determine that if an accountability system rewards only test scores, then the easiest way to look good is to find a school with a record of high achievement, and that frequently means running away from the problems of poverty, hunger and violence--and denying leadership in schools where it is most needed.

  In fact, several urban schools are demonstrating they can achieve high standards even in the presence of seemingly intractable poverty. Nevertheless, this success takes years to achieve, and it is a constant challenge in the face of mobility, poverty and a host of factors that interfere with the ability of students to learn. Leaders who accept this challenge deserve our gratitude rather than a litany of complaints about why they can't make schools effective.

  How can an accountability system reconcile the tension between the

demand for achievement and the need to recognize incremental improvement? The comprehensive accountability system in the Milwaukee Public Schools has won wide recognition for its balanced approach. It is a three-tiered system, including systemwide goals, school-based goals and a school narrative.

The six systemwide goals focus on student achievement and attendance. Using state and local test data, the board established thresholds that are expected of all students in all schools. The five school-based goals are selected from more than 100 goals that focus on specific instructional strategies at the school level. Each faculty, site council and principal collaboratively selects the five school-based goals, and the accountability system provides credit based solely on improvement. This allows the lowest-performing schools to achieve recognition for progress, even in the early stages of an improvement program when the systemwide goals are not yet achieved.

In the same way, the focus on improvement challenges the highest-performing schools by requiring them to find areas of improvement, even if they already meet the systemwide targets. Thus every school is challenged, and every school finds opportunity for reward.

The third tier of the comprehensive accountability system, the school narrative, allows principals to explain the meaning behind the numbers--the context, school climate and professional environment. Taken as a whole, this system is a model for providing clear targets for principals while allowing variation among complex and widely varying schools. The system also enhances communication with the media and the public and makes the essential case that accountability is more than test scores.

Simple But Particular

- *Communicate specific demands.*
  Several states and a growing number of school districts have concluded that effective accountability for school leaders

79

requires linking school performance to the job security of the principal. In this way of thinking, when students underperform, the administrator's certificate is on the line.

Even if one accepts the dubious argument that optimal performance results from threats and intimidation, those on the receiving end of the threat must understand what they are supposed to do. Unfortunately, the complexity of some accountability systems borders on the absurd, and one recent study estimated than only a fraction of the principals who had been warned that their jobs were on the line understood how their activities influenced the results of the complex regression equations driving the accountability system.

- *Focus on student learning.*
Telling principals to focus on student achievement represents a "blinding flash of the obvious," to use Tom Peters' well-turned phrase. However apparent such a focus may be, the real life of a principal is dominated by anything but student achievement. Parent meetings, discipline and a seemingly unending series of paperwork and meeting demands all take time away from serious consideration of student achievement.

At the secondary level particularly, the expectation that a principal will observe more than a few moments of each class is unrealistic, yet the principal is expected to make decisions of enormous importance on tenure and promotion.

It doesn't have to be this way. Stan Scheer, superintendent in the Ferguson-Florissant, Mo., School District, follows a tradition of putting himself on the substitute teacher list for 20 days each year--a practice he started as a principal. When he is on the floor with a group of kindergarten students for several hours, he undoubtedly is missing some calls from school board members and meetings with the chamber of commerce. He also is sending an unambiguous message about the value he places on student achievement and the daily lives of students and teachers. In essence, he places his calendar where his rhetoric is, and that is an act all too rare.

How else can principals demonstrate a

80

focus on achievement? Three specific strategies are used in effective schools.

First, principals themselves can establish the meaning of "proficiency" in an academic area. Flora Flagg, a Milwaukee principal, creates a writing prompt for her 4th-grade students every week and personally applies the citywide writing rubric to these papers. It is no surprise that the students in this 100 percent poverty-level school scored in the 74th percentile on national writing tests.

Second, principals can personally review a sample of student portfolios. Another Milwaukee principal reported that when she reviewed just three portfolios per month, teachers reported a significant increase in the quantity and quality of student portfolio contents.

Third, principals themselves can monitor the performance of students who face the possibility of failure. With the same intensity that we follow the performance and injury statistics for sports stars on a weekly basis, we can track the most critical information on selected students. Principals cannot be expected to personally attend to the performance of hundreds of students. But principals can follow the performance of those who risk failure on a "watch list."

When these three steps are taken, principals uniformly report two important results--students have several significant adults who consistently care about them, and principals feel less like chief disciplinarians and administrators and more like instructional leaders.

Periodic Feedback

- *Ensure frequency of assessment, evaluation and feedback.*
  I am stunned constantly by the number of principals and senior administrators who tell me they have not received formal evaluations during the past three years. Some have never been evaluated but received periodic raises, encouragement, threats or pink slips. This erratic feedback usually came from board members and superintendents who extolled the virtues of accountability but could not find the time to practice it

81

themselves.

Even those superintendents who conscientiously provide annual feedback to principals should not be complacent. Annual appraisals encourage a binary approach to principal evaluation: success or failure, met the goal or missed the goal, keep your job or look for a job.

If feedback is to be effective, it must be frequent. We know how ineffective annual feedback is for students and teachers. They receive annual test scores (typically several months after taking the test and in most cases long after the students have departed for the summer). For teachers, such feedback sends a message that "here's the data that would have helped you improve your teaching based on the needs of these students if you would have had it in time, but since it's late and there's nothing you can do about it, we'll just release it to the newspaper so that they can editorialize again about how bad our schools are."

Most educational leaders would agree this practice is hardly a prescription for motivating teachers; yet we commit the same error when we provide feedback to principals on an annual or less frequent basis.

The most effective feedback systems provide measurable results at least quarterly. If the results are to be taken seriously, they are carefully considered. This review takes time, and that is something few superintendents have. Therefore, evaluations must be focused and brief so that, within a 10-minute meeting and a two-page memo, the key points of principal evaluation can be addressed.

Some districts evaluate schools and principals by publicly posting evaluations outside of each school. This is fine, provided the evaluation data can change frequently. This sends the message that "we may not be where we want to be yet, but we're getting better every day." Unfortunately, the more common approach is to place a sign outside of a school based on the results of the previous year and then leave the sign up all year. This sends the message, "we came here labeled as a failure, and we're stuck here until next year, so don't

82

expect things to get any better, because
we're going to remind you every day you
walk in this building just how bad things
are."

- *Adapt to individual strengths.*
  John Goodlad is among the leading
  exponents of the importance of school
  leadership and the difference a single
  leader can make. Yet even Goodlad
  rejects the notion that the principal
  should be fiscal manager, chief
  administrator, personnel director,
  disciplinarian, curriculum designer and
  instructional leader. "It is naïve and
  arrogant," he writes, "to assume that
  principals, who may or may not have
  been effective teachers, can acquire and
  maintain a higher level of teaching
  expertise than teachers engaged in
  teaching as a full-time occupation.

  "The concept becomes particularly
  absurd at the secondary level, where
  presumably the principal who has
  attended some special institutes on
  teaching, necessarily for short periods of
  time, will have acquired teaching
  competence beyond that of the teachers
  of each of the diverse subjects ... I
  certainly would not want to put myself in
  such a posture of universal excellence!"

  We do expect a great deal of principals,
  but the best leaders know above all their
  own weaknesses. Their response is not
  camouflage but team-building, seeking
  leadership team members whose
  strengths complement those of the
  principal. An effective evaluation system
  explicitly recognizes that leaders need
  not be superb in every category but
  rather encourages them to achieve
  results with teams rather than follow the
  ineffectual and mythical model of the
  Lone Ranger.

Beyond Amusement

- *Reward the tough choices.*
  We often confuse effectiveness with
  popularity and reward the leader who is
  better at self-promotion than instructional
  leadership. This occurs predominantly in
  curriculum choices and staff
  development.

  Consider the extraordinary state of

curriculum options in secondary schools. I recently visited a high school in a relatively small community. Both the district and the state had established academic content standards, and administrators spoke eloquently about the need to focus on the academic needs of an astonishingly large number of underachieving students. I asked, "With the majority of your students failing 9th-grade algebra, would you consider providing additional math instruction in middle school and high school so these students would have a better chance on the state graduation exam?"

"Oh no, we can't do that," came the quick reply. "Our parents, students and teachers really like all of our electives--in fact, we have more than 400 different courses in this high school, so we just don't have time for the 9th graders to take more math."

The next few years in this district are predictable. Students will perform miserably on the state exam, the principal will be replaced and the superintendent will be fired. Their successors will join the Rotary Club and make some rousing speeches--but they had better not touch the unfocused curriculum that got the district into trouble in the first place.

Staff development poses a similar challenge for leaders who must reconcile popularity with effectiveness. The "Dr. Fox" research shows our frequent response to teaching effectiveness. Adult students listened to an entertaining but vacuous presentation by Dr. Fox and raved about his insight and the educational and professional value of the experience.

Alas, Dr. Fox was a professional actor who provided a series of empty platitudes and encouragement, but none of the professionals in the audience saw through him. Worse yet, the real expert who followed Dr. Fox with solid content and essential information was barely tolerated by the audience. This is not to say that staff development need be boring or dominated by lectures. However, it must be more than entertainment, and leaders must reject demands for programs that are distinguished more by amusement than

84

substance.

Of course, tough choices are not limited to curriculum and staff development. Bus schedules, cafeteria food, coaching changes, library books, student rules and rights and a host of other decisions test the effectiveness of school leaders. The popular model of seeking input in all decisions has its limits, particularly if the standard for the final decision is what is best for student achievement rather than what is most popular.

- *Take time for reflection and self-evaluation.*
  Teachers and parents know that among their most important jobs is the emancipation of the student from the need for minute-to-minute guidance. Only when the student, confronted with a difficult decision, can reason out a sound solution independently will the parent and teacher be able to say that their hard work has paid off.

  School leadership involves hundreds of independent decisions, yet many leaders are evaluated on the basis of compliance rather than judgment. An effective evaluation system can help principals distinguish between these two criteria by requiring the person being evaluated to write a brief reflection and self-evaluation.

  Among the many inaccurate caricatures of Asian schools is the notion that schools are filled by automatons who never exercise independent judgment. In fact, an integral part of the educational ethic in Japan (and several other Asian countries) is the notion of hansei, or self-evaluation and reflection. The Third International Math and Science Study astonished many observers when it noted that Japanese math classes included debates and discussions of alternative answers—a far cry from the Western stereotype of stale recitation in the Asian classroom.

  Leadership evaluation can benefit from a healthy dose of hansei by beginning every evaluation conference with a serious self-reflection on what the educational indicators for that school district imply about leadership, decision making and future strategies. This self-reflection should be revealing to the

85

superintendent and board. Is self-evaluation a series of excuses, a litany of achievements or a serious attempt to learn from the past and construct a better future?

No one doubts that accountability by school leaders is imperative. The goal of improved student achievement is best achieved not by intimidation and threats, but by a system that focuses on academic improvement while providing specific and meaningful feedback to students, teachers and leaders.

Doug Reeves is president of the Center for Performance Assessment, 1660 South Albion, Suite 1110, Denver, Colo. 80222. E-mail: TestDoctor@aol.com

For Further Reading
Doug Reeves suggests the following works for those who want to learn more about accountability for principals:

"Accountability in Complex Educational Systems," by Douglas B. Reeves, available from Center for Performance Assessment, 800-844-6599

Holding Schools Accountable: Performance-Based Reform in Education, edited by Helen F. Ladd, available from the Brookings Institution, 202-797-6252

Transforming America's Schools: An Administrators' Call to Action, by John Murphy and Jeff Schiller, available from Open Court Publications, 800-852-0790

A Place Called School: Prospects for the Future, by John I. Goodlad, available from McGraw-Hill, 800-999-6430

http://www.aasa.org/publications/sa/1998_10/Reeves.htm

7/1/2005