Lessons from a Comprehensive Performance Appraisal Project

A five-year effort to develop a model evaluation program for school systems in Iowa and Minnesota has yielded valuable information for others.

Suppose you had enough money, just this once, to develop a *complete* system of teacher and administrator performance evaluation, one that would spend additional time and resources to fill the gap every time someone said, "Yes, but what about evaluating _____, which is omitted?" What would such a total-systems approach to evaluating and improving the performance of educational professionals look like? What does "total" mean? How would such a system evolve and operate? Perhaps most important, how much would it cost to operate?

Shirley Stow and I, as codirectors of the School Improvement Model (SIM), funded in large part by the Northwest Area Foundation of St. Paul, had a once-in-a-lifetime opportunity to find out. Beginning with a research planning grant in 1979, we were to spend five years putting together a model that has as its cornerstone measuring and improving how administrators lead and teachers teach to affect how children learn. Now that the final report to the foundation is completed, this is the first article dealing with the various components of the model (see fig. 1).

In the School Improvement Model, everyone receives feedback from everyone else: teachers from students, principals from teachers, superintendents from principals, and researchers from stakeholders.

Photograph by Mark Lasoff

Background

The SIM consortium included the Minneapolis Public Schools, Northfield Public Schools, Edina Public Schools, and Breck School (Independent), all in Minnesota, and Spirit Lake Community Schools in Iowa. Planning for the project began in 1978, and the experiment spanned the school years 1979–80 through 1983–84. Our goal was to demonstrate the effect of an articulated system of administrator and teacher performance appraisal (with interventions to encourage productive behaviors) on pupil achievement in mathematics and reading at the elementary and secondary school levels as measured by both norm- and criterion-referenced tests. We operationally defined quality of educational program in terms of student achievement.

Our colleagues at the Research Institute for Studies in Education (RISE) gradually established a division of labor that resulted in some specialization.[1] All of us contributed to performance appraisal. Working with stakeholders' committees in each school organization, representing what Bruce Joyce (1986) would call "responsible parties," the RISE researchers labored to produce five separate appraisal systems that were valid, reliable, and would be legally discriminating (i.e., would separate high performance from low). We used a three-year process in each school organization: one year to plan, one year to field-test with volunteer teachers and all administrators, and a third year to fully implement the system with all educational professionals. Thus, our intent was to relate performance appraisal, supervision, and staff development.

While school culture accounted for some differences, we noted several dominant themes in the planning year. The particular time-window of the project, 1978–1984, no doubt influenced the perspectives of the participants. Committees felt the pressure of events such as the urban crisis in education, a press for accountability, declining enrollments, the conservative backlash, greying-but-staying faculties, funding cutbacks by state governments, and a need to involve teachers' unions in school reform efforts. A few of the project's slogans sketch the philosophical premises we adopted.

- "Teachers are the solution, not the problem."
- "Start with board and administrator performance evaluation because good bosses set good examples."
- "People don't do what you expect; they do what you inspect."
- "What gets measured gets accomplished."
- "If you want it taught, test it!"

Don't jump to the conclusion that we took a heavy-handed, sweatshop approach to performance appraisal. Teachers were the dominant group on each stakeholder committee, and, in general, their standards were higher than those of other stakeholders—a fact that may surprise some who don't listen to teachers! We did, however, keep everyone's attention focused on student learner outcomes. Moreover, we stressed that effective classrooms are nested in effective schools, and we disaggregated student achievement scores by gender, race, and socioeconomic status. We provided training to all participants to enhance their performance, instead of merely "evaluating." Finally, we never let anyone off the "high-expectations hook." Everyone received feedback from everyone else: teachers from students, principals from teachers, superintendents from principals, RISE researchers from stakeholders, and outside trainers, such as Sam Kerman for TESA (Kerman et al. 1980), from pre- and post-testing of trainees.

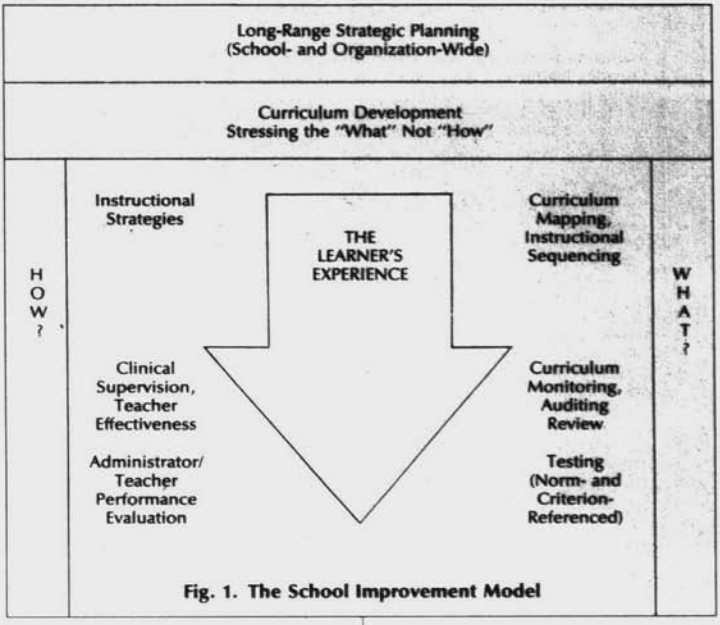

Fig. 1. The School Improvement Model

The Yes-Buts

The project's research team spent a great deal of time listening to stakeholders, entire faculties, and outside experts. Each time we thought the system was complete, someone would say, "Yes, what you have accomplished is good, but what about . . . ?" Shoring up each of these weak areas in our plan proved costly, but informative. Besides, that was the mission of the School Improvement Model, namely, to try out many high-risk activities with venture capital from nonschool sources in order to build a model that would serve both public and independent school organizations. Many consultants, trainers, and educational researchers advised us during the five years. Those who contributed the most important "yes-buts" were Valerie Broughton, Gordon Cawelti, Ron Edmonds, Tom Good, Madeline Hunter, Sam Kerman, Larry Lezotte, Tom Romberg, and Ernie Stachowski.

Starting with a rather straightforward notion of improving student

1. Maintains an effective relationship with students' families.
2. Provides instruction appropriate for capabilities, rates of learning styles of students.
3. Prepares appropriate evaluation activities.
4. Communicates effectively with students.
5. Monitors seatwork closely.
6. Demonstrates sensitivity in relating to students.
7. Promotes positive self-concept in students.
8. Promotes students' self-discipline and responsibility.
9. Uses a variety of teaching techniques.
10. Spends time at the beginning of the learning demonstrating processes to the student (cueing).
11. Uses controlled (guided) practice before assigning homework (independent practice).
12. Organizes students for effective instruction.
13. Provides students with specific evaluative feedback.
14. Selects and uses appropriate lesson content, learning activities, and materials.
15. Demonstrates ability to monitor student behavior.
16. Writes effective lesson plans.
17. Demonstrates a willingness to keep curriculum and instructional practices current.
18. Has high expectations.
19. Organizes resources and materials for effective instruction.
20. Models and gives concrete examples.

This list is a composite of discriminating criteria used by one or more of the five SIM school organizations.

Fig. 2. Discriminating Teacher Performance Criteria in Rank Order

"We strongly embraced the contemporary definition of an effective school—one in which student achievement cannot be predicted by socioeconomic status, race, or gender."

learning by evaluating and improving how teachers teach, we added the following elements:

- long-range strategic planning at the district and building level;
- curriculum mapping to enhance alignment and density;
- criterion- and norm-referenced testing twice a year, which offered diagnostic information at the beginning of the year and gain scores by class and building at the end of the year;
- board performance evaluation, which provided a potent example;
- administrator performance evaluation to "evaluate the evaluator";
- student ratings of teachers, which were more discriminating than any other source of teacher data once proper instruments were developed;
- school climate surveys to measure such important contextual variables as faculty cohesiveness, goal orientation, and expectations for student achievement;
- professional improvement commitments—what George Redfern (1980) would have called "job improvement targets"—to provide written agreements between the teacher and the evaluator for improved performance in the next cycle;
- protocol materials in the form of videotaped classroom segments with teacher work samples to be judged and rating norms to train appraisers and measure reliability;
- student demographic data such as gender, race, and socioeconomic status;
- attendance data for both teachers and students;
- training for teachers composed of classroom management, essential elements of effective instruction, and Teacher Expectations for Student Achievement (TESA); and
- a statistical design to relate all project data, for which we chose a causal model called Path Analysis.

Essential Questions

Stakeholders' committees in each school organization sought to answer four questions when setting up formative and summative evaluation procedures for teachers and administrators.

1. What are your performance evaluation criteria? (How would you know an effective teacher if you saw one?)

2. How high are your standards? (The mere presence of behaviors isn't enough; the real question is how appropriate are they?)

3. How will you monitor and report performance? (Teacher observation is an established procedure, but how do you monitor an assistant superintendent for finance?)

4. How do you help an evaluatee improve after you have established a profile of strengths and weaknesses? (Professional improvement commitments are effective, but they are very difficult to write in the middle of a rather negative summative evaluation conference.)

In retrospect, each committee's answers were remarkably similar, even though each spent a year in deliberation and every group of "responsible parties" represented organizations solicited for educational diversity.

Criteria

Stakeholders and their university-based consultants sought valid, reliable, and discriminating performance criteria. Validity meant that the criteria measured what they purported to measure: good teaching or effective principal behaviors. Reliability meant consistency (i.e., that multiple evaluators agreed on what the teacher/administrator was or was not doing well). Discrimination power was established experimentally but meant simply that the criteria must reveal real differences in performances. Those experienced with faculty-wide performance data know too well the frustration of discovering that every teacher was evaluated "superior" and every principal rated "excellent"!

SIM researchers have written extensively about the criteria selection process elsewhere (Manatt et al. 1976, Stow and Sweeney 1981, Manatt 1985, Look and Manatt 1983, Manatt and Stow 1986). It suffices here to say that criteria for teacher performance evaluation were validated by effective teaching; principal criteria were validated by school effectiveness research; and criteria for central office administrators were validated by task analysis, job descriptions, and critical work activity time-motion-study.

As Ned Gage (1983) so aptly puts it, "Research on teacher effect provides suggestions about how a teacher should behave on a continuum from a hunch to a trace to an imperative to a categorical imperative." Some of the evidence is so strong that for a teacher to omit certain behaviors probably constitutes malpractice. SIM researchers feel the same way about the principal and central office administrator criteria. Nevertheless, we were able to produce a powerful set of criteria for elementary and secondary principals and a less potent, but very usable, set of evaluative instruments for 64 central office positions. To establish reliability in the field, we had multiple

1. **Sets instructional strategies/emphasizes achievement**
   - **Promotes activities to identify, analyze, and solve instructional problems.**
   - **Emphasizes student achievement with teachers and students on a regular basis.**
   - **Has high expectations for student academic achievement.**
2. **Supports teachers**
   - **Organizes a system in which teachers work cooperatively to develop and implement instructional objectives.**
   - **Encourages a free and open flow of comments, suggestions, and recommendations.**
3. **Coordinates instructional program**
   - **Defines goals and objectives of the school and works toward articulation between schools and grades.**
   - **Monitors the curriculum and identifies progress toward stated curriculum/program goals.**
4. **Provides orderly atmosphere**
   - **Schedules instructional space for maximum use and strives for minimum disruption of instruction.**
   - **Sets high standards of conduct and monitors all facets of school life to ensure that these standards are met.**
5. **Promotes professional growth**
   - **Provides support and direction for those staff members seeking to improve their skills.**
   - **Makes regular, systematic, and cooperative appraisals of each staff member's performance, always including a follow-up conference.**

This partial list is for illustrative purposes only.

Fig. 3. Discriminating Performance Criteria (Secondary Principals)

GRADES K-2
1. My teacher gives us enough time to do our work.
2. I pay attention in class.
3. My teacher takes a lot of time before starting teaching.
4. I understand the lesson being taught.

GRADES 3–6
1. My teacher gives us work to do at home.
2. I can get help from my teacher.
3. I finish my work before class is over.
4. My teacher has us work too fast.

GRADES 7–9
1. My teacher gives homework related to the subject we are studying.
2. My teacher makes classwork interesting.
3. We use one book at all times in the class.
4. My teacher often makes materials and worksheets for us to use.

GRADES 10–12
1. Some students disrupt or bother the class when we are working.
2. My teacher asks questions to see if we understand what has been taught.
3. When we finish a lesson, we discuss and summarize what we have just studied.
4. My teacher often loses his or her temper when students disrupt class.

This partial list is for illustrative purposes only.

Fig. 4. Student Rating Items

appraisers rate the same evaluatee, using a long list of criteria. At different times and in many and differing school organizations, we have had students, peers, principals, and other administrators rate teachers and principals. All of the sample criteria in Figures 2 and 3 are discriminating at the .01 or .05 levels.

Reliability was established for the criteria by field ratings of large numbers of teachers and administrators. Reliability on the Kuder-Richardson-20 ranged from .597 for primary students rating teachers to .80 for principals rating teachers to .91 for central office administrators rating principals (fig. 4).

Standards

We used a Behaviorally Anchored Rating Scale (BARS) to express standards of performance. This scale's response mode describes the extent to which an evaluatee exhibits the behavior or performs the criterion. Such a response mode (fig. 5) increases the likelihood of scatter (namely, spreading out the ratings so that all do not receive a superior rating), as well as solving the problem of defining "incompetent teaching." None of the 50 states has an adequate legal definition of incompetence, but all recognize a school organization's right to maintain standards of work performance. Thus a teacher or principal in a SIM school organization might have been evaluated as performing below "district standards" but would never be called "incompetent." This proved to be a very important distinction later when we set up programs of intensive assistance for marginal teachers and, when necessary, dismissed a subpar performer.

Monitoring and Reporting

Too many teacher evaluators are like angels; they make visitations not visits. That is, their physical presence in the classroom significantly alters what goes on. This is especially true when the principal or other evaluator is in the classroom so seldom that it becomes an event. We discovered a rather surprising tendency, namely, that typical teacher evaluators overestimate by 100 percent the amount of time they spend evaluating one teacher. Surveyed at the start of the project, evaluators estimated 12 hours per teacher. When we actually timelogged the evaluation process, the time total for one teacher turned out to be six hours. This time allotment discrepancy gets intertwined with the persistent lament, "I don't have time to evaluate all my teachers," which seems more socially acceptable to say than "I won't give it enough time" or "I don't know how."

To answer the monitoring-and-reporting question for both administrator and teacher evaluation, it is first necessary to differentiate formative and summative evaluation. Formative evaluation is ongoing, nonjudgmental coaching and counseling, which is done to improve teacher performance. Summative evaluation is judgmental and comparative, and perhaps, if the teacher is subpar, adjudicative. Summative evaluation helps management make better decisions. Management in this context extends all the way up to the governor and the state legislature. Formative evaluation is done for accountability.[2]

SIM researchers discovered that summative evaluation can be done in six to eight hours, but to really improve teacher performance via formative evaluation requires much more time. Indeed, the best teacher evaluator in our sample, a devotee of the UCLA teacher decision-making model (Hunter 1984) averaged 29 hours per teacher per year in teacher evaluation activities. Judging from the five-year experience, we recommend two announced visits with a pre- and postobservation conference and informal drop-by visits as often as possible. These are minimums, and the best principals will devote much more time. To properly evaluate principals or division heads, we found that superintendents and headmasters must spend at least as much time.

Reporting performance is generally less scientific and less accurate than monitoring. Imagine for a moment that you own the largest new car dealership in your community. Can you envision evaluating the 1986 performance of your sales manager without including some measure of how many cars the sales force moved last year? Would you use only a checklist to see if he or she looked good? Of course not. *Money* is at stake.

In schools, student learning is the bottom line, but we found that superintendents, headmasters, and other individuals who evaluated principals never looked at how teachers performed as a part of the principal evaluation process. Indeed, they couldn't because teacher performance data

were stored in the buildings with no comparisons made across departments, buildings, and the entire school organization.

We solved that problem first by creating a computer-based system of comparisons that required a mainframe computer to do the analyses, and later by developing Computer Assisted Teacher Performance Evaluation/Supervision (CATE/S) for microcomputers. CATE/S provides performance reports by evaluator, department, building, and the district. It also generates reports of staff development needs and suggests professional improvement commitments for weak teachers and professional growth plans for superior teachers.

Getting Better

Strategies to help teachers and administrators improve their performance are the crux of the School Improvement Model project. We call them professional improvement commitments (PICs) or professional growth plans. Because such plans are better developed away from the stress of evaluation conferences, the research team worked with a group of highly skilled clinical supervisors to write several PICs for each commonly used teacher performance criteria. Next we critiqued the PICs with a panel of experts, assembled them in a compendium, and gave a copy to each evaluator. The use of teacher PICs is now rather routine, but principal PICs are being created at this writing (fall 1986). Once the PICs were field-tested, it was a fairly simple task to hook them to teacher performance profiles via microcomputers using CATE/S programming. Anyone who might think it is "unprofessional" for a skilled principal to use CATE/S should know that medical doctors have CADUCEUS, which routinely outperforms internists in diagnosing and prescribing for patients.

Moreover, we noted that evaluators can be sloppy with a manual system of evaluation, but a computer-based system (especially one that uses scanform input) encourages accuracy and promptness.

Lessons

The School Improvement Model used performance appraisal of educational professionals as the centerpiece for a total-systems approach to school improvement. We started with the not-yet-accepted idea that "a set of school practices shown to promote learning in one school can do the same in any school." Further, we strongly embraced the contemporary definition of an effective school—one in which student achievement cannot be predicted by socioeconomic status, race, or gender. Basic to the entire effort was the philosophic premise that three levels of observation are necessary: classrooms, schools, and the macroview of the entire school organization.

We learned many lessons in such a broad and lengthy experiment. Here I will enumerate only those closely linked to performance appraisal, supervision, and staff development.

1. Administrator evaluation is not a difficult process once criteria and procedures are established; however, cabinet-level officials are often reluctant to devote enough time to the task to do it properly.

2. Teacher performance evaluation is a complicated but necessary process. School improvement is contingent upon changing how teachers perform. Teacher evaluation is particularly complex because teachers and styles of learning and teaching are so varied. To be sure it is done well requires a special kind of principal evaluation.

3. Developmental/participative supervision for teachers is a difficult change for principals to make. Consequently, principal performance evaluation must be linked to the motivation matrix for principals (i.e., to make it happen you must change how principals "keep score").

4. A "people change" is more important than a "paper change." We found the following characteristics to be essential in a positive evaluation relationship:

a. Performance criteria must make sense to teachers and administrators (we linked them to research on effective teaching and effective schools). Employee rules must be recognized as important, but separate.

b. Cooperative efforts must be developed between evaluatee and evaluator.

c. Evaluator and evaluatee must communicate honestly and forthrightly.

d. Participants must be sensitive to each other's concerns and responsibilities.

e. Objectivity and clearly delineated expectations are essential. Indeed, we taught principals, when describing teacher performance, "If you didn't see it, it didn't happen; and you didn't see it if you didn't write it down!"

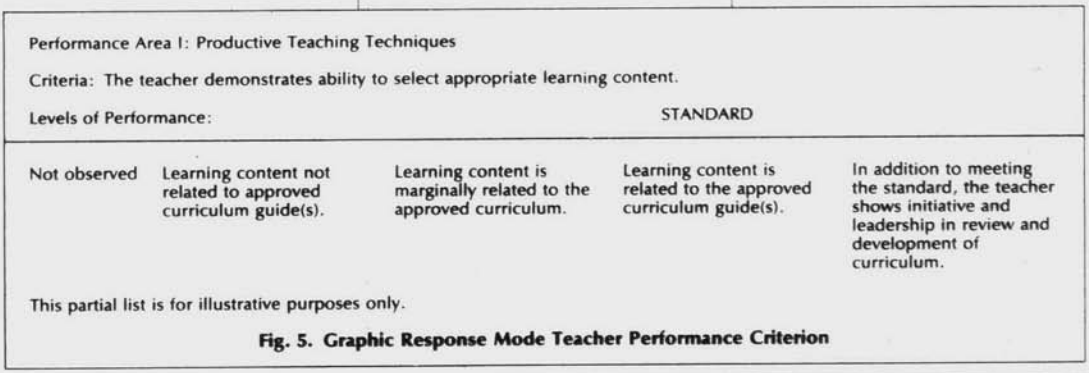

Fig. 5. Graphic Response Mode Teacher Performance Criterion

f. The goal should not be weeding out a few subpar performers but rather improving the performance of all teachers.

g. Both evaluator and evaluatee must practice confidentiality.

All educational professional evaluators learned that we evaluate because we believe that every professional is capable of improving his or her performance. The probability that improvement will occur increases when evaluation is carried out systematically and in accordance with good planning.

5. Rater bias and lack of knowledge of effective teaching research make it difficult to prepare skillful teacher evaluators. Our validity and reliability checks indicate that teacher volunteers can become effective teacher evaluators faster than seasoned principals.

6. Analysis of teacher performance data across departments, buildings, and the entire organization is crucial to planning staff development programs that really help teachers improve performance. Such analyses are not difficult to do and result in more *usable* information for superintendents, headmasters, and boards.

7. Job targets or professional growth plans are essential to effective appraisal. Left untrained for this task, principals will simply approve "work activities" that teachers were going to do anyway. This mistake occurred less frequently in administrator evaluation, perhaps because management by objective is more established among school managers.

8. TESA, essential elements of instruction, and classroom management training were the most commonly used staff development interventions. In general, they had a neutral effect on student achievement.

9. Teacher and student attendance has a considerable impact on student achievement. Five to seven days' absence appears deleterious for students. A slightly longer period of absence for teachers, say seven to ten days, is significant. Attendance almost always seems to be an important factor in students' achievement in mathematics, less so in reading. Investigators conducting applied or theoretical research experiments simply must pay more attention to how much time intervention training causes teachers to spend away from students.

"Perhaps the biggest surprise in the total-system approach is the significant effect of pre- and post-testing (with proper reports to teachers) on student achievement."

10. All subordinates want to "evaluate the boss." We found that measuring school climate affords a prime opportunity for teachers to give principals feedback on what teachers expect from them and what they perceive principals are giving them.

11. Perhaps the biggest surprise in the total-systems approach is the significant effect of pre- and post-testing (with proper reports to teachers) on student achievement. The advantage of twice-a-year testing with criterion-referenced measures is so great that the SIM research team will stress this link extensively in the next series of experiments.

12. Finally, how much does performance appraisal cost? Using "ahead of the art" benefit-cost-analysis techniques, we found that teacher performance evaluations cost from $116 to $242 per teacher per year across four of the SIM school organizations. Even Breck School, which used a multiple evaluation team to determine merit pay, spent only $216 per teacher per year. Twice-a-year norm- and criterion-referenced testing cost $5 per pupil per year. Of course, in the classrooms where achievement increased, it didn't cost—it *paid*!□

1. J. Stanley Ahmann and Richard Warren were responsible for research design; Tony Netusil for board performance evaluation; Jim Sweeney, school climate measures; Shirley Stow, performance improvement commitments; and Charles Ruebling and Ann Thompson, student achievement testing.

2. When the crisis in urban education reduced confidence in our schools, governors (and their legislative colleagues) took the initiative away from school administrators and began mandating new evaluation approaches and career ladders; 27 states had done so by spring 1986. See T. F. Allen, "Identifying Behaviors of the Master Teacher" (doctoral diss., Iowa State University, Ames, 1986).

References

Gage, N. L. "When Does Research on Teaching Yield Implications for Practice?" *Elementary School Journal* 83 (March 1983): 492–496.

Hunter, M. "Lesson Design." In *Using What We Know About Teaching*, edited by Philip Hosford. Alexandria, Va.: Association for Supervision and Curriculum Development, 1984 Yearbook, 121–141.

Joyce, B. *Improving American Schools*. New York: Longman, 1986.

Kerman, S., T. Kimball, and M. Martin. *Teacher Expectations and Student Achievement Coordinator Manual*. Bloomington, Ind.: Phi Delta Kappa, 1980.

Look, E., and R. P. Manatt. "Performance Criteria for the Evaluation of School Principals and Headmasters." Occasional Paper No. 83–4, Iowa State University, Ames, June 1983.

Manatt, R. P. "Competent Evaluators of Teaching: Their Knowledge, Skills and Attitudes." In *Practical Applications of Research*, a CDER monograph edited by W. R. Duckett. Bloomington, Ind.: Phi Delta Kappa, 1985.

Manatt, R. P., K. Palmer, and E. Hidlebaugh. "Evaluating Teacher Performance with Improved Rating Scales." *NASSP Bulletin* 60, 401 (September 1976): 21–24.

Manatt, R. P., and S. Stow. *Developing and Testing a Model for Measuring and Improving Educational Outcomes of K-12 Schools: Technical Report*. Ames: Iowa State University, February 1986.

Redfern, B. *Evaluating Teachers and Administrators: A Performance Objective Approach*. Boulder, Colo.: Westview Press, Inc., 1980.

Stow, S. B., and J. E. Sweeney. "Developing a Teacher Performance Evaluation System." *Educational Leadership* 38 (April 1981): 538–541.

Richard P. Manatt is Professor of Education and Director of the School Improvement Model (SIM) for the Research Institute for Studies in Education, Iowa State University, Ames, IA 50011.