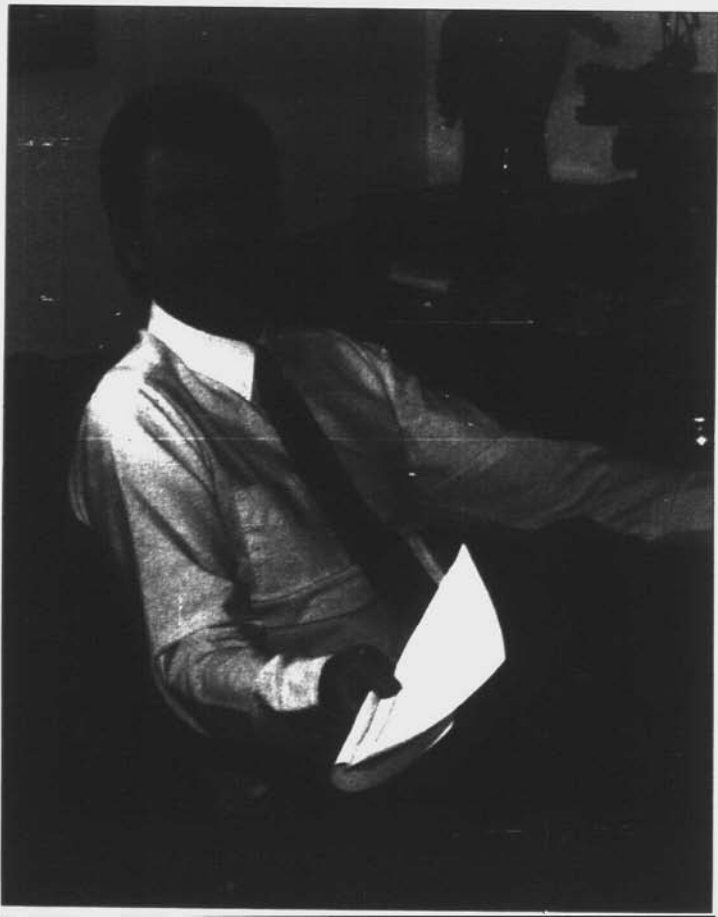# On Teacher Evaluation: A Conversation with Tom McGreal

Tom McGreal, Chair, Department of Administration, Higher and Continuing Education, University of Illinois, Urbana-Champaign, has consulted with numerous states and school districts on design of staff evaluation programs and is the author of ASCD's *Successful Teacher Evaluation* (1983).

You've often said that the only instrument you need for teacher evaluation is a blank sheet of paper. Do you still believe that?

If we had complete freedom to choose, and well-trained evaluators, yes, but there've been a lot of forces working against this in the last five years. We've been making considerable progress in the use of instructional improvement-oriented goal-setting models. But unfortunately, many of the states—around 30 at this point—have mandated certain forms of teacher evaluation that more often than not require districts to submit plans for state approval that are more accountability oriented. As a result there is renewed attention being given to a rating scale mentality. So even though a lot of us don't like that approach, local districts are asking for help in improving the quality of rating scales.

In what way?

Well, the most significant trend in teacher evaluation in recent years has been the heavy emphasis on the use of the research on teaching as a focus for the criteria.

Rather than a more general "trait" approach?

Yes. We're seeing a tremendous amount of similarity in the criteria that districts and states are coming up with. That's partly because they are sharing criteria, and partly it's a reflection of the limited number of criteria you can generate from the research we now have on effective teaching.

Is there a danger that, because these criteria can only reflect the current state of research, they will be too narrow?

I think that's a clear problem. Like many people who work in this area, I feel fairly positive toward using the research on teaching, but there are many cautions about using it. As everyone surely knows, the criteria that have been developed are much more appropriate for explicit instruction than for other forms of teaching. In a number of the states where I've been working, we've begun to see that instruments using the research on teaching criteria tend to give an advantage to certain teachers. For example, we're seeing a high percentage of special education teachers and mathematics and science teachers showing up very well, perhaps because their training and the content they teach fit more readily, while other teachers—particularly those who try to stress higher-level thinking skills—are put at a disadvantage. So we have to be very careful about use of these criteria, especially if they're tied to incentive plans.

If you have these concerns, there must be some good reasons for going ahead and using the effectiveness criteria anyhow.

There are basic teaching skills that every teacher ought to be able to demonstrate. I think the effective teaching findings have great potential, but they must be used with understanding, and I'm not sure that understanding always exists.

What do people need to understand?

Evaluators—and teachers themselves—need to have some perspective on where the research on teaching comes from, what it says, and what it doesn't say. It's also important to understand the differences between the teacher effects research and the Madeline Hunter material and its derivatives. States, particularly in the South, are mandating that every teacher be trained in the Hunter-derived approaches, such as the Principles of Effective Teaching (PET) program. PET was developed, I think, by Don Roberts and others when Roberts was superintendent at Newport News. He took it to Arkansas when he became

"We can discriminate a competent from an incompetent teacher, and we can usually distinguish extraordinary teaching from the rest, but anything beyond that is virtually impossible with any degree of reliability."

Commissioner of Education there. It's now moving into Texas and South Carolina and a number of other states.

How do you differentiate between the Hunter material and the effective teaching research?

There're obviously some similarities, but they're really two different strands. Local schools and states need to understand that the effects research is a combination of correlational and experimental studies tying individual teacher behaviors to student outcomes. The Hunter material and the programs derived from it are learning theory-based, and are often viewed as directed at the total teaching act. Hunter's is a complete teaching approach as opposed to some of the more visible teacher effects programs such as TESA (Teacher Expectations and Student Achievement) and the Evertson and Emmer (1981) work that focus on certain aspects of teaching. There is little evidence at this point to link the Hunter training to student learning. Now, that's not necessarily a criticism of it, but it is a criticism of states and school districts that promise that it's going to improve student learning. We just don't have much evidence for that. What's more, although Hunter and most responsible Hunter advocates try to keep their training separate from evaluation, there's no question that state department people and local administrators are taking the Hunter terminology and recommendations and turning them directly into rating scales.

You said you have some qualms about the use of rating scales, but you've been helping states and school systems develop better ones anyhow. What are some ways to make rating scales as good as possible?

I might explain that most of us who specialize in evaluation don't have a lot of faith in most rating scales, but the reality is that they are going to continue to be prevalent despite our reservations, so help should be given to make them as useful as possible. One way districts are trying to make their instruments more objective is by using "behaviorally anchored" ratings: each rating has a brief description of what it means. For example, a satisfactory rating might mean "teacher shows clear evidence of having established goals and objectives." This does add a certain amount of reliability to a rating scale.

The process actually results in a numerical score, then?

Many of the states and local districts expect evaluators to give an overall rating. So their systems are set up with a formula that derives an overall rating from the ratings on individual criteria. For example, in order to get an overall rating of "exceeds district standards," you have to get "exceeds district standards" on at least 9 of 13 criteria and no lower than "almost meets district standards" on each of the other 4. Well, that's tough to do. Among other things, it assumes that all 13 criteria are equally important, which is almost certainly not the case. I recommend not giving an overall rating but, if there must be one, having only two categories: "meets district standards" or "doesn't meet district standards."

If that's not acceptable, I'll try to get them at least to weight the criteria. Some can be viewed as more fundamental than others, and the overall rating should reflect that.

Another thing we need to do is to incorporate what we know about teaching for higher-level learning in evaluation criteria. Right now, other than a few vague references like "uses a variety of questions," such provisions are virtually nonexistent.

As I mentioned earlier, one thing I suggest is to look for ways to weight criteria that are better at discriminating at the lower ends of rating scales versus those that might discriminate better at upper ends. Our ability to

evaluate performance is, at best, relatively gross. We can discriminate a competent from an incompetent teacher, and we can usually distinguish extraordinary teaching from the rest, but anything beyond that is virtually impossible with any degree of reliability. So I suggest that criteria like "use of time," "bell-to-bell planning," and "teaching to objectives" be considered fundamental expectations used to discriminate between competent and incompetent or satisfactory and unsatisfactory teaching. Other criteria, such as the teacher's ability to match methods with content and with kids, are much higher-level skills. They should probably be used to discriminate between "satisfactory" and "superior" or between "meets" and "exceeds" district standards, and so on. The purpose is to take some of the pressure off the principals and supervisors who are forced by the system to make discriminating judgments among teachers, and then are provided no guidelines or help in responding to teacher questions like, "What do I need to do to receive a 'superior' rating rather than an 'excellent' rating?" If we were to ask the effects researchers to weight the various criteria, I am sure we would not find

them receptive to this notion. I can understand their reservations since the research was not designed to do this. On the other hand, I think that it's reasonable to suggest that certain teaching variables are more fundamental and should be displayed if a teacher is to receive a minimally acceptable rating, while other behaviors, though desirable, cannot be required to meet minimum expectations. I think we owe local supervisors this if we are to help them cope with the demands that their evaluation systems put on them.

You do believe, then, that an instrument can be helpful in discriminating between the competent and the incompetent?

Yes.

I suppose most members of the general public think that's the main purpose of teacher evaluation.

That's true. And they don't see why it should be difficult, especially when we have rating scales, because we're supposedly getting objective data. But the use of rating scales is actually very subjective.

A fundamental problem is that almost all of our supervisory training

deals with formative data collection: how to write descriptively rather than judgmentally, techniques of clinical supervision, conferencing skills, and so on. So we now have a group of administrators being required to take formatively collected data and make summative judgments from them. Making that transition is very difficult, if not impossible.

A long-standing concern is making the same individual responsible for both, and having teachers never know for sure just what the administrator is up to at any moment.

The evaluation system can be a big help with that. Otherwise the whole system becomes rating-scale driven. It forces supervisors to make one or two observations a year using a wide-angle lens perspective because they have an instrument in front of them which they end up using as an observation guide. The conference inevitably focuses on the judgments the supervisor has already made, and he or she never gets a chance to use the skills he or she has been trained to use.

The system should allow a supervisor to—at least artificially—separate his or her administrative behavior from his or her supervisory behavior. I have found that an individual goal-setting model can help supervisors do this. You have monitoring of performance going on all the time. Teachers aren't evaluated once a year or twice a year; they're evaluated continuously—because they live and work in the same place with their administrators every day. I try to set up a system in which the monitoring of minimum expectations is continuous. Then, less frequently, maybe every other year, the supervisor and teacher sit down and identify one or two fairly specific teaching goals that become the basis for what they will work on together. That puts them in a much better position to use formative techniques. They can focus their observations more clearly; they can use clinical models; they can establish a training, coaching environment.

Using the process you've described, there still comes a time when a teacher may have to be dealt with in a different way.

Right. Poor teaching is still going to show up. In goal setting, the rule tends to be that the less competent or mature the teacher, the more directive the supervisor has to be. The more competent and mature the teacher, the more he or she should be involved in the activities and setting of the goals. Since the majority of teachers in this country are mature and competent, it seems sensible to design the system for them: for the roughly 98 percent who are going to be in that district for a long time.

Do you advocate some kind of special notice for the other 2 percent? Something like, "I am hereby informing you that you are now in a different category. Here's how it will work"?

Yes, we probably have to do that. It keeps the special mechanisms that apply only to the marginal teacher separate from the regular, more formatively oriented system.

But you're also saying that it's difficult to develop and maintain a goal-setting model when you have to comply with mandated systems that require everybody to be treated as though they're on notice all the time.

Yes, state departments need help in somehow walking that line. Now, I don't want to pick on state departments; many times they're required by law to put together rules and regulations for statewide mandates. But it's of particular concern that a lot of these programs—whether they require use of a single system or they force local districts to conform to certain kinds of requirements in order to get approved—take away the flexibility of local districts to build systems that are more conducive to improving instruction.

The other side of the argument, I suppose, is that these laws wouldn't have been necessary if local administrators had been more responsible all along. A little earlier you said that 98 percent of teachers are doing a good job. A lot of people feel that's probably an overstatement; that considerably more than 2 percent need some attention.

"A fundamental problem is that almost all of our supervisory training deals with formative data collection: . . . So we have a group of administrators being required to take formatively collected data and make summative judgments from them."

In some cases that's correct. The state must insist that school districts that are doing nothing—and believe me, there are some—be brought up to minimum standards. The tendency in recent years, unfortunately, has also been to bring districts that have already seriously addressed the issue back down to the middle.

But to return to your original question, when I say that 98 percent of teachers are doing their jobs, I'm not saying they're all doing the job as well as they could. I'm saying they're doing the job to the point where it would be virtually impossible to have them dismissed; that, because of the political realities and the cost of dismissing teachers, a purely summative evaluation system will probably not affect their behavior very much.

If that's the case, what is a better strategy?

The important thing is relating evaluation to staff development. If teachers feel more involved in setting their own goals, they'll be more committed to improvement.

I thought you said you were concerned about the danger of using the Hunter model in both staff

development and evaluation, but now you're saying the two *should* be tied together.

Let me explain. Researchers have helped identify a set of skills that ought to be in every teacher's bag of tools—certain organization and management skills, skills in the use of time, Rosenshine's (1986) teaching functions. We have clearer information about how to improve the quality of seatwork and practice activities. We have a much clearer sense, particularly for junior and senior high school teachers, of what constitutes a climate conducive to learning. We know more about increasing levels of student involvement, opportunities for success, appropriate use of feedback, and so on.

A first stage of appropriate staff development would be to give all teachers and administrators an introduction to this bag of tools. From there you can move to more complex topics, such as learning styles and thinking skills, but we need to have in common this relatively sound base.

Now, these tools are there for teachers to sort through and think about, not always to have to use. One benefit of having a set of fundamentals is that it provides a common language teachers and supervisors can use to talk together. We find that it increases the level of talk about teaching, and that's true of the Hunter material as well.

There are appropriate ways to link this kind of staff development to the evaluation system?

Right. Eventually you change your evaluation system to reflect the staff development program. In many ways the staff development and the evaluation system are subsets of something bigger; maybe it's a five-year commitment to enhance the quality of classroom instruction (I'm purposely using the term "enhance" rather than "improve" because it connotes working with people who are already fairly competent). Still later the evaluation system becomes the means to maintain the staff development. In many states and districts where there's been good staff development over several years, we've seen considerable erosion when the original training is discontinued and staff turnover begins to take its toll. When that happens, partic-

ularly if the district is moving toward a goal-setting model, the only mechanism to maintain the staff development is the evaluation system.

You're saying that, even where a district must follow state mandates, it may be appropriate to start with staff development on effective teaching, develop some shared understanding among teachers and supervisors, and then gradually introduce a goal-setting approach when the staff is ready to handle it?

That's exactly what I'm trying to encourage. Even though districts are being required to use rating scales and so on, I'm saying they can still keep alive on an every-other-year basis, or if necessary an every-third-year basis, a more intensive involvement between the supervisor and the teacher.

One interesting development in teacher evaluation is participation by teachers in the evaluation of other teachers. In some cases excellent teachers are taken out of the classrooms for a year or more to spend full-time acting as evaluators.

That sort of thing is going on in a few exemplary places, but it doesn't tend to be transferable to other settings. Those of us who spend a lot of time improving skills of administrators and supervisors are seeing a growing movement toward mentoring and peer coaching. Now I think there's a tremendous need for more interaction among peers, but in the schools I've worked in, the average teacher still does not want to be involved in evaluating other teachers. So while there are extraordinary examples of peer supervision, I think there are many of us who feel that for the near future the real hope of improving instruction is still going to come from improving the interaction between the supervisor and the teacher. I think it is wrong to move toward a model that forces administrators to be responsible only for summative evaluations and asks teachers to make formative evaluations. By providing appropriate staff develop-

ment opportunities and designing evaluation systems that encourage and allow good supervision, we can establish quality supervisor-teacher interaction that can enhance classroom instruction. We have the knowledge and ability to do both.☐

References

Emmer, E. T., and C. M. Evertson. "Synthesis of Research in Classroom Management." *Educational Leadership* 38 (January 1981): 342–347.

McGreal, T. *Successful Teacher Evaluation*. Alexandria, Va.: Association for Supervision and Curriculum Development, 1983.

Rosenshine, B. V. "Synthesis of Research on Explicit Teaching." *Educational Leadership* 43 (April 1986): 60–69.

Tom McGreal is Acting Chairperson and Associate Professor, Department of Administration, Higher, and Continuing Education, University of Illinois at Urbana-Champaign, 1310 S. Sixth St., Champaign, Il. 61820. **Ron Brandt** is ASCD's Executive Editor.