Pitfalls and Pratfalls of Teacher Evaluation

W. JAMES POPHAM*

Isolated here are certain errors to be avoided by would-be teacher evaluators. One should review these difficulties "for even in teacher evaluation it is true that the sins of the fathers/mothers are often visited upon their sons/daughters. And who needs that kind of a visit?"

ANTHROPOLOGISTS have amply documented the proposition that there are certain perduring albeit fluctuating interests of a society. At widely separated points in the society's development these dormant interests emerge quite prominently, only to return soon to a state of extended hibernation. In our society, for example, we can safely predict the resurgence this decade of the yoyo or the hula hoop. And the man who does not save his narrow (or wide) neckties during the offseasons is either affluent or foolish.

In American educational research we can document a similar phenomenon with respect to the interest in teacher effectiveness research. Since the beginning of the twentieth century we have witnessed a series of sporadic interest peaks in the assessment of teacher competence. An examination of past educational research annual meeting programs or educational research journal archives will reveal occasional flurries of activity regarding this important topic. And there is little question that today we are flat in the middle of one such flurry. A perusal of current educational research association programs will reveal a considerably larger proportion of papers and symposia dealing with teacher effectiveness questions than was the case just a few years ago. For the moment, at least, prediction studies of college grade point average are out and teacher effectiveness investigations are decisively in.

Future historians of education may derive special delights from deciding whether today's current interest in teacher effectiveness stems more directly from our impending teacher surplus (We now have the luxury of being selective, hence require selection indicators.); from the current concerns regarding educational accountability (California's

* *W. James Popham, Professor of Education, University of California, Los Angeles*

A few years ago when someone attempted to assemble all of the popular classroom observation schemes it took not one, but two large volumes to do the job. Once more we must be cautious not to equate popularity of usage with defensibility of usage.

recently approved teacher evaluation law requires annual teacher assessment.); or from some other yet ill-defined cause (for example, climatic or astrological variables). For educational researchers whose interest in teacher effectiveness assessment has been recently kindled, or rekindled, the lessons of history are well worth mastering, for we may thereby avoid a few of the myriad cesspools into which our predecessors have tumbled.

The remainder of this analysis will consist of the identification of the more prominent perils which have plagued teacher effectiveness researchers and which promise to prove equally vexing to that recently emerging but rapidly multiplying species—the *teacher evaluator*. Visitors to Hollywood are often delighted to find that the tourist can secure, at reasonable cost, a map of movie stars' homes. Abetted by such maps, the curious midwesterner can peer with adulation only at the homes of those Hollywood greats who have personal significance to him. He can avoid those mansions in which stars of lesser luster reside, or homes of those whose films he despised.

Perhaps teacher effectiveness researchers and evaluators can view this analysis in a similar vein. Perhaps by anticipating some of the more prominent errors of the past we can help today's noble folk avoid prevalent pitfalls. Perhaps they can employ this paper as their map of menacing muckups. For a muckup, once isolated, can be more skillfully sidestepped. Let's examine a few of these traps.

Sinning From Scratch

The initial trap to be avoided by teacher evaluators stems from an all too common dismissal of history. In most respectable scientific endeavors the creator of today stands, metaphorically, on the shoulders of giants, that is, those whose prior contributions make possible current advances. Even though the field of teacher effectiveness assessment is not surfeited with our predecessors' breakthroughs, it would be a grave mistake to adopt a tabula rasa stance with respect to this area of inquiry. There are a number of excellent reviews of teacher competence research available in such sources as the *Encyclopedia of Educational Research* and the two *Handbooks of Research on Teaching*. If for no other reason than to discover which assessment strategies proved fruitless, a careful review of prior work is mandatory for any first rate teacher evaluator.

Teacher Effectiveness Research vs. Teacher Evaluation

Educational researchers who have been reared on admonitions regarding the dangers of overgeneralization should be particularly sensitive to the proposition that findings from teacher effectiveness research investigations should not be blithely transferred to a teacher evaluation framework. Yet, some of our most revered methodologists have suggested that because researchers have discovered certain teacher behaviors which tend to be related to learner growth, such teacher behaviors can be used as the chief indicators by which a particular teacher can be evaluated. This is bad advice.

For several reasons this particular recommendation is faulty, yet the overriding error is that when an individual teacher is being subjected to a personal evaluation, an evaluation which may seriously affect that teacher's livelihood, a totally different situation is created than when that same teacher participates in a far less threatening investigation designed to "isolate instructional variables of relevance to learner growth." In one instance a teacher is allowing data to be

gathered for a scientific (often a doctoral dissertation) mission. In the other instance the stakes (for the teacher) are incredibly higher. The research and the evaluation settings are so dissimilar that in many instances generalizations from one to the other are not warranted.

Perseverating on Perverse Rating

Rating teachers is easy. It is popular. It has been going on for so long that we assume it has merit. But, by and large, it doesn't.

Many educators believe that if a rater has enough opportunity to observe a teacher in action, then it is possible to produce an overall or multidimensional rating of that teacher's instructional capability. Yet to date there has been scant evidence that such ratings are sufficiently well correlated with pupil growth to warrant their widespread use. The difficulty with the use of a rating strategy for evaluating teachers is the criteria employed by the raters. They are characteristically so ill-defined and so variable across raters that the results of evaluator ratings are typically inconsistent and confusing. Different people value different things in teachers. Different raters have different perceptions of good teaching. Such diversity leads to grave difficulties in evaluating teachers via ratings.

If no practical alternatives to a rating strategy exist, then ratings are probably better than nothing—particularly if they are employed only to isolate the extremely weak or extremely strong teachers. But those employing such rating schemes should not be deluded into thinking that because ratings are widely employed they should be widely applauded.

Observe, If You Will . . .

Right behind ratings in this week's teacher evaluation popularity poll we find systematic observations. During the past decade the use of observational schemes has really burgeoned, with Ned Flanders and his Interaction Analysis Acolytes leading the action. A few years ago when someone attempted to assemble all of the popular classroom observation schemes it took not one, but two large volumes to do the job. Once more we must be cautious not to equate popularity of usage with defensibility of usage, for there are real problems associated with employing observation schemes for teacher evaluation.

Perhaps most prominent is the fact that classroom observation schemes typically focus on the teacher's behavior or, at best, the interactions between teachers and pupils. Clearly, such a focus is on *process* criteria (what goes on during an instructional sequence) rather than *product* criteria (what happens to learners as a result of the instructional sequence). Now when we evaluate a particular teacher's capability we cannot assume that the presence of process criteria will inevitably lead to the occurrence of desired product criteria. In spite of a given teacher's use of desirable instructional procedures, there may be aversive procedures, personality quirks, etc., which although not included in the observational scheme (an *exhaustive* observational system is clearly impractical) work to cancel out the positive features. For evaluating a given teacher, and that is what teacher evaluators are about, observational systems carry this key liability.

A second difficulty with observation systems is that these days most self-respecting and legalistically moxie teachers' organizations will require any teacher evaluation enterprise to spell out its evaluative criteria in advance—thereby inviting teacher fakeability during observed classroom sessions.

Seduction of the Standardized

In spite of increasingly frequent structures against the use of standardized tests for evaluating the quality of instructional programs or teachers, there are still those researchers who believe that if they can find a test listed in one of Buros' *Mental Measurements Yearbooks* then such a measuring instrument, so anointed, can be used with confidence in almost any research or evaluation situation.

But for purposes of teacher evaluation

it is generally conceded that standardized tests possess far too many drawbacks. The main problem with such tests is that their chief function is to permit comparisons among individual *learners*, not *teachers*, hence they possess properties which are almost antithetical to teacher evaluation. For instance, because of the requirement to produce considerable score variance, some oft-revised standardized tests end up with a raft of achievement items most closely correlated with the learner's native intellectual ability. Items of this sort are relatively impervious to the effects of even high quality instruction.

Another serious limitation of standardized tests is that, because they are usually designed to service an entire nation, their content coverage is extremely wide. Often the curricular emphases of a teacher or school will not be consonant with those of the standardized test, thus a misleading mismatch is created between the instructional treatment and the measurement device.

The Lure of the Logical

We must be wary of the seductive appeal of logic—more precisely of logically derived techniques for dealing with problems that of necessity require empirically substantiated solutions. Some teacher evaluators have become so enamored with the conceptual carpets they weave that they are loathe to see how well they work. And some of these carpets will fray badly if walked on by a single spider—or a married one.

The writer pleads guilty on this count, for when conjuring up the notion of teaching performance tests a few years ago, he viewed the logical appeal of the approach as overwhelming. Why not see how effectively teachers could accomplish *given* instructional objectives during short lessons (as reflected by learners' post-lesson performance), then generalize regarding that teacher's ability to accomplish his/her own objectives in more typical settings? By providing identical objectives to teachers and assigning learners randomly, comparisons could be made across teachers.

Now all of this may be conceptually alluring, but until a ton of hard data is at hand to demonstrate that teaching performance tests are reliable, valid, and practical, we must withhold unequivocal support of the approach. Teacher evaluation will surely be an evidence-based game, and our evidence-gathering strategies must themselves withstand the scrutiny of both logical analyses and evidence-based appraisals.

Enchantment With New Solutions

It is not surprising that people who have been frequently frustrated will turn with fervor to anything that can possibly alleviate their problems. In teacher effectiveness assessment we have wandered down so many blind alleys that any new path may look appealing. There have been those educators so entranced with their discovery of Interaction Analysis as an observation-evaluation system that they were prepared to make all of their observations on stone tablets and to canonize Ned Flanders soon thereafter.

One new-found convert to teaching performance tests proclaimed with inordinate pride at a recent professional meeting that he had devised "the first reliable measure of general teaching ability." Most of his time at the meeting was spent putting copies of his "breakthrough" papers in the hands of anyone who would take them.

Quite clearly, moderation is called for in appraising the value of new assessment schemes. Just because prior procedures have not proved successful, we may tend to grasp at any new straw we see. But even a brand new straw may be incapable of bearing the weight demands imposed by current teacher evaluation requirements.

Unstable Standards of Scrutiny

Everyone has favorites. And most teacher evaluators have preferences regarding assessment strategies. We viscerally feel that certain approaches ought to work. But prudence dictates that we view supporting evidence for all assessment schemes with the same standards of rigor, not relenting in the case of the preferred approach.

In a recent analysis, one of our most respected research methodologists spent page after page flaying one approach to the assessment of teacher competence. Then, having rejected the despised approach on the basis of an intensive study-by-study refutation of the strategy, advocated an alternative approach largely on the basis of second-hand reviews of the pertinent investigations. Such variable standards of rigor may be useful for promoting controversy. They are inappropriate for solving as tough a problem as how to evaluate teachers.

Legislation Sans Legislature

There will very likely be an increasing number of states that enact laws similar to California's 1971 Stull Act which requires the annual evaluation of teachers. Those individuals attempting to implement laws established by properly constituted legislatures should give an ethical re-think when they advocate implementation schemes clearly at variance with legislative intent.

In California, for example, though the Stull Act specifically requires the teacher to be evaluated in part on the basis of evidence of student growth, some implementers have advocated a total reliance on classroom observation or pupil rating schemes.

Another part of the law required that teachers be appraised with respect to their "adjunct duties" such as extracurricular assignments. Yet some implementers have argued that one of the truly professional teacher's adjunct duties should be self-evaluation, hence the adjunct duty aspects of the evaluation system can be concerned exclusively with a teacher's self-evaluation efforts.

Now if teacher evaluators believe that legislative enactments are unworkable or otherwise ill-advised, they have an ethical responsibility to protest against the legislation. They should not tamper with a law once it leaves the legislative chambers. Subverting the intent of legally constituted lawmakers should not be one of the teacher evaluator's competencies.

The Mr. Chips Pitfall

By and large most teachers are human. Most humans have frailties. Many of these frailties are based on a person's perceived need to survive and prosper. We should not be surprised, therefore, if teachers do not submit docilely to evaluation enterprises, but instead try to beat the game. Not every teacher is a Mr./Ms. Chips who thinks exclusively of the good of the little folk.

This recognition is not intended to denigrate teachers. In general the vast majority of teachers are trying to do the best job they can for children. But when their job is on the line because of a teacher evaluation operation, we should expect that teachers will behave in a human, not saintly manner. For instance, if they are to be evaluated in terms of achieving certain instructional objectives, then we must expect that some teachers will eschew all hard-to-attain objectives, and go for the sure winners. If they are to be observed, we must expect that some will cleave to desired practice during the observations. In other words, the astute teacher evaluator will have to recognize that all teachers are not benevolent reincarnations of Socrates. Nor are they malevolent monsters. They're just people, and people-evaluation schemes must incorporate procedures for coping with human frailties.

The Top Ten

In review, an attempt has been made to isolate certain errors to be avoided by teacher evaluators. The attentive reader will have noted that there are precisely ten of these errors, hence one might conceive of Biblical parallels replete with a list of "Thou shalt nots. . . ."

But while no nightly litany of error avoidance is being advocated, it may be useful for teacher evaluators to review the difficulties noted here. For even in teacher evaluation it is true that the sins of the fathers/mothers are often visited upon their sons/daughters. And who needs that kind of a visit? □