George E. Flower

Will-o'-the-Wisp Assessments of Teaching Performance

A running controversy, which pops up periodically from province to province and school system to school system, centres on the possibility (or impossibility) of relating teachers' salaries to quality of performance. We have not heard the end of it, by far; for it seems clear that, with teachers constituting a steadily growing proportion of Canada's work force (one in forty-two as of 1959, one in thirty-five as of 1964), the time will come when it will not be possible for the level of salaries to rise nearly high enough to attract and retain truly outstanding teachers. Indeed, we lose many such teachers now; as exceptionally good teachers they are promoted out of teaching and into administration, for our salary schedules do provide differentials for increased responsibility. Yet, what more vital responsibility is there in a school than the direct responsibility of the teacher to help youngsters learn?

The difficulty of relating teachers' salaries to quality of performance, of course, lies in obtaining and agreeing upon valid and reliable measures of that performance. Most school systems which have attempted so-called merit plans have, in fact, simply added something which boils down to a long-service bonus, or a differential based on the teachers' participation in curriculum development work, or community service, or some other such thing, not related to any direct assessment of relative effectiveness in the classroom. (In passing it is interesting to note that, while recognition of outstanding teaching in salary schedules is not generally acceptable to teachers, recognition of outstanding teaching, by the same unsatisfactory measuring sticks, *is* acceptable for promotion out of the classroom!)

What is generally accepted, by teachers and administrators and public alike, is that good teaching is essential for an effective school program. This means that teaching is being evaluated and must be evaluated constantly, quite apart from any possible relationship to salary. Such evaluation is essential to serve two crucial purposes.

The first purpose is the administrative one of providing some rational basis for necessary personnel decisions. Should this teacher's contract be renewed, or should he be counselled out of teaching? Which of our teachers should be pressed into service as a consultant in language arts, or science, or whatever? Is the total staff of a school relatively weak, and in need of strengthening? And what shall the school administrator reply to the Dean of Education at McGill when he inquires about the work of Mr. X or Miss Y who is being considered for appointment to the staff of the Faculty of Education? Personnel records for teachers have consisted too often merely of their initial letters of application, vital statistics as to successive teaching assignments and salary changes, plus a note or two of lateness or absence in the winter of 1964-65. There is clearly a need, for purposes of personnel decisions, for systematic evaluation of the work of every teacher and for recording that evaluation.

By far the more important purpose of evaluation of a teacher's performance, however, is to locate his strengths and weaknesses and help him to recognize them, so that the one may be reinforced and the other improved. This is the truly creative reason for evaluation: so that every teacher may be helped and stimulated to perform more effectively.

How shall a school system arrange for such assessments of teaching performance to be made? A variety of approaches are followed, but a common procedure, for better or for worse, is for a principal or inspector to observe a teacher at work and to record opinions based on his observations. The logic behind this approach is that presumably a major qualification of the principal or inspector is that he has the ability to distinguish between good and less-than-good teaching. It is his very knowledge and expertness in appraising teaching and learning that sets the *school* official apart from other types of official. We would expect a school principal or inspector or education officer to be far more accurate and helpful in his evaluation of the work of a teacher than would be, say, a bank manager or an architect or a naval officer.

But how good are ratings by school administrators? Can they really tell good teaching when they see it? (Breathes there the school administrator — or teacher, for that matter — who has not said to himself, at least once, "Give me half an hour with a teacher at work, and I'll come up with a pretty shrewd opinion as to whether that work is good, bad, or indifferent"?) An obvious way to put this to the test is to have a group of school administrators observe the same performance by the same teacher, and see if they arrive at similar ratings. While this is difficult to arrange in a live classroom, it is easy to arrange with television or a filmed

performance, such as the kinescopes developed for this purpose in a large-scale research project titled the Development of Criteria of Success in School Administration.[1]

I have used one of these kinescopes, showing Miss Walenski teaching a twenty-minute lesson in arithmetic to her grade one class, with various groups of principals and inspectors over recent years. After the "visit," each participant was asked to complete an evaluation sheet outlining strengths and weaknesses as he saw them, and to assign a global rating of Miss Walenski's performance on a five-point scale. How did the 392 individual ratings compare? Results were as follows:

Excellent	—	3%
Better than average	—	28%
Average	—	41%
Doubtful	—	27%
Very weak	—	1%

Not only was there a wide difference of opinion among these supposedly capable judges, but the spread was so great as to take on much of the appearance of the normal curve. My colleague Professor J. Glenn Scott has had similar results with this and other films. Worth of Alberta used the same film and a seven-point scale with a group of sixty-five principals, again with substantially similar results.[2] He also checked to see whether more experienced principals tended to vary less in their ratings than did less experienced principals, but found no significant difference. In a controlled situation, then, when each rater has precisely the same information or lack of information about Miss Walenski and her class, and is asked to appraise precisely the same performance, the resulting overall appraisals vary widely.

Why?

When faced with that question, many of my 392 appraisers suggested that they had been asked to formulate a judgment on the basis of inadequate information; twenty minutes in Miss Walenski's room was not enough. True. Yet it can be argued that one rarely if ever has all the information one should have before having to make a decision. The participants also pointed out that they were, and should be, less interested in global evaluations than in specifics. If evaluation is to be a guide for future action to help Miss Walenski improve her performance, it does not help very much to label her as "better than average" or "doubtful." When a person visits a physician for a check-up, he expects some specific advice, and not a report that "Your general state of health is three on a five-point scale." This is particularly important with reference

to the second purpose for evaluating a teacher's performance: namely, to assist and motivate that teacher to greater effectiveness. Nonetheless this begs the question. Global evaluations are being made, and they differ widely.

Many factors would appear to contribute to this variation in assessment, including the following:

1. Differences in the observers themselves; in the accuracy of their perception, in their preoccupation as to what is important and what is not, in their expectations for the particular situation, in their previous experience with teachers in similar situations.
2. Lack of agreement among observers as to desirable goals of education and desirable teaching-learning processes.
3. The many variables in any teaching situation, which make it difficult to generalize as to what is effective and what is not. Thus Teacher A teaches X to Pupil B. Every teacher is different from every other teacher, and every pupil is different from every other pupil. Perhaps the "X" (that which is to be taught and learned) can be more clearly specified; yet even here variations will have to enter in, if the strengths of the particular teacher are to be utilized to meet the needs of the particular pupil.[3]

In the face of all these difficulties, how can will-o'-the-wisp assessments of teaching performance be made more valid and reliable? There have been many studies in this most elusive of fields. At one point the "critical incident technique" looked promising. Why not have competent judges describe briefly, in behavioural terms, some critical incident — some snippet of teacher behaviour they have observed recently — that made them think, "There is excellent teaching." Similarly, have them describe a critical incident which they judged indicative of very poor teaching. Gather thousands of such reported incidents, analyse and classify them, in the hope of ending up with a list of behaviours typically involved in or critical of excellent teaching, and typically involved in or critical of poor teaching. But as the New England School Development Council discovered, the results are disappointing.[4] For one thing, an incident reported by one observer as descriptive of outstanding teaching will be reported by another as precisely the opposite.

More promising work has been done in the way of developing more elaborate and systematic guides for observing and appraising the work of teachers. A good example is Ryan's Classroom Observation Record, which consists of twenty-two items developed on the

basis of available research.[5] Final ratings may be derived on three major clusters of observable teacher behaviours:

1. understanding, friendly *vs.* aloof, egocentric, restricted
2. responsible, business-like, systematic *vs.* evading, unplanned, slipshod
3. stimulating, imaginative, surgent or enthusiastic *vs.* dull, routine.

Such forms, however, especially when they become very detailed, can be used in mechanical and meaningless ways. They are probably most useful when it is realized that (a) they do not simply *add up* to a global evaluation, but they do provide somewhat more precise data for judgment as to what that global evaluation should be; and (b) they are best used diagnostically, to direct the attention of appraiser and teacher alike to particular aspects of the teacher's performance.

Further, surely the ultimate criterion of effective teaching is growth on the part of the pupil toward desirable educational goals. In judging teacher competence, then, great emphasis should be placed on the pupils and their progress, rather than upon details of teacher behaviour *per se*. It is suggested that, if we really want to be more certain of our appraisals of teaching performance, we had better devote a great deal more attention to being clearer as to what desirable educational goals for pupils are, and to our measurements of pupil progress toward those goals. (It should be noted that while we shudder at the thought of judging teacher competence mainly by pupil results on examinations, at the same time we are reasonably content to have pupils pass or fail, be promoted or not promoted, mainly on the basis of those same results on examinations. While I would not suggest that we should therefore evaluate teaching more than we already do on the criterion of pupils' scores on examinations, that would be just as defensible as not paying more attention to other factors in our evaluation of pupils' performance.)

Andrews, among others, has listed various ways for improving global ratings of teaching performance.[6] The following appear to me to be promising approaches:

1. Use ratings by as many competent judges as possible, including more than one rating by the same judge, and average the ratings.
2. Ensure that each rater's concept of good teaching is in keeping with the school or school system's official statement of desirable goals of learning.
3. Have raters look for evidence of pupil accomplishment such as improvement over a period of time, attitude to work,

curiosity, rather than concentrate only on the current process of teaching.

4. Have raters realize that good teaching may appear in many different styles; what works for one teacher may not work for another.

5. Have raters compare their ratings with those of others, so that they may learn to recognize and compensate for their own biases.

Finally, if it be accepted that the overriding purpose of evaluating teaching performance is the improvement of instruction, then it follows that all such evaluations should be made known to and discussed with the teacher. It is the teacher, after all, who is going to have to behave differently if his performance is to be more effective. The process of evaluation is only beginning when an appraiser arrives at an opinion, assigns a rating, lists strengths and weaknesses. The process ought to continue to include consultation with the teacher, discussion with the teacher, follow-up with the teacher, decision by the teacher as to what he can begin to do to improve his performance, however excellent it may already be.

Notes

1. See John K. Hemphill, *et al., Administrative Performance and Personality,* New York: Bureau of Publications, Teachers College, Columbia University, 1962.

2. Walter H. Worth, "Can Administrators Rate Teachers?" *The Canadian Administrator,* 1:1 (October, 1961).

3. See Jane R. Martin, "Can There Be Universally Applicable Criteria of Good Teaching?" *Harvard Educational Review,* 33:4 (Fall, 1963), pp. 484-90.

4. David V. Tiedeman, ed., *Teacher Competence and Its Relation to Salary,* Cambridge: New England School Development Council, 1956.

5. David G. Ryans, *Characteristics of Teachers: Their Description, Comparison and Appraisal,* Washington: American Council on Education, 1960, pp. 86-92.

6. John H. M. Andrews, "The Evaluation of Teaching Service." Address to the 1963 Convention of the Canadian Education Association. Reproduced (mimeo.) by the Association. See also Ryans, *op. cit.*