DOCUMENT RESUME

ED 300 882 EA 020 239

AUTHOR Scriven, Michael
TITLE Evaluating Teachers as Professionals.
PUB DATE 13 Mar 88
NOTE 42p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; Administrator Evaluation; Elementary Secondary Education; Negotiation Agreements; *Teacher Evaluation; *Teacher Qualifications; *Teaching (Occupation)
IDENTIFIERS *Professionalism

ABSTRACT

This document outlines a practical teacher evaluation system that avoids the fatal invalidities of present methods. The recommended approach treats teachers as responsible professionals undertaking to perform certain duties while retaining considerable autonomy in discharging them. While teachers acknowledge a need for accountability and systematic professional development, they also deserve full protection against the use of an arbitrary, invalid, unjust, or noninformative system. The discussion first explains why commonly used approaches are invalid, highlighting teaching as a profession, the limits of negotiation (or political compromise), administrator accountability, the lack of serious administrator evaluation systems, professional development connections, summative evaluation versus development support, and the need for a comprehensive evaluation system. The paper then shows the underlying fallacies of research- and judgment-based tests and of management-by-objectives approaches. The duties-based model is presented as a valid alternative that divides teacher merit into four major categories and considers a teacher's worth to the district. Finally, an exhaustive list of minimally required teachers' duties is presented and the new rules and sources of evidence explained. The inclusion of student test data, the teacher portfolio containing self-evaluations and personal development plans, and "footprint" data (exit interviews with graduating seniors) more fully depict a given teacher's contribution. (MLH)

EVALUATING TEACHERS AS PROFESSIONALS

Michael Scriven

University of Western Australia[1]

TABLE OF CONTENTS

[1] Nedlands, Western Australia, 6009. The author can also be reached at his California address: POBox 69, Pt. Reyes, CA 94956

I: PRELIMINARIES

1. Overview

This document outlines a practical system for teacher evaluation that avoids the fatal invalidities of present methods. Most of the components are tried and true; the novelty lies in the exclusion of invalid elements and the inclusion of enough valid ones to provide an adequate foundation for personnel decisions[1]. The approach recommended is based on treating teachers as responsible professionals who undertake to perform a roster of professional duties but retain a great deal of autonomy about the means they use in discharging those duties. This implies that they acknowledge the need for accountability and systematic professional development, each of which requires an evaluation process (the processes overlap considerably, but are not identical). But it also implies that they should expect, and get, full protection against the use of a system that is arbitrary, invalid, unnecessarily intrusive, unjust, or unable to provide information of use to them. The main discussion begins by laying out the reasons for concluding that alternative approaches *are* invalid. Unless this point is convincingly established, there is no reason to get involved in the effort of setting up a new system. In the next few sections, we address a number of background, conceptual and organizational issues that need to be handled before any specific procedures can be set up[2]. These sections provide definitions and clarifications—and even procedures—that are crucial to the main section.

2. Teaching as a Profession

No assumption is made here that teaching is or is not at the moment and as a whole, a profession in the full sense of the term. It is clear enough that many teachers behave like professionals and that many others do not, a situation not unlike that in other professions. However, teachers as a whole have never set up the codes of professional conduct or boards of ethical review that characterize the more highly-regarded professions. They often teach in schools which represent their clients' only way to get their children educated, while it is almost always possible for clients of lawyers and doctors to get second opinions, a process which provides part of a system of checks and balances. Teachers are also relatively immune from consumer suits for incompetence. And it is rare for teacher organizations to announce full responsibility for updating their repertoire of knowledge and skills without employer support (beyond identification of the deficiency), as is more common in the better-

---

[1] These include appointment, tenure, retention, promotion, merit increases, advantageous transfers, awards, sabbatical leaves, and the opposite of each.

[2] This is a discussion of the evaluation of teachers in the primary and secondary schools, but essentially similar principles apply to tertiary teachers; the major difference is the weighting of the non-classroom obligation to do research.

paid professions. In fact, many teacher unions have explicitly rejected this responsibility. So it can be argued that teacher organizations are in a situation where they ought to be highly accountable, yet have shown little sign of accountability.

But individual teachers and some of their subject-matter associations accept a general responsibility for keeping up-to-date on content and to some extent on pedagogy, one of the most critical features of the professions. And there is no doubt that, like the typical professional, teachers work in an environment in which many of the obligations of the job are understood rather than spelled out in job descriptions. In fact, it turns out that a model of evaluation appropriate to the professions provides the best 'fit' for the evaluation of teaching. Indeed, it is argued here that it is the only model that is both feasible and valid. It thus seems clear that teaching is best *conceived of* as a profession, whatever the proportion of teachers that rise to or reject that conception. The approach laid out below, or something close to it, is arguably an essential part of a blueprint for progressing towards full professional status.

Teachers have to consider both advantages and disadvantages about being treated as professionals. Representatives of industrial unions often deride the 'professionalism' thesis as a way to get more work out of teachers without any more pay. An alternative and often-voiced view is that there is very little chance of teachers getting significantly better pay scales—in many districts—without some corresponding increase in professionalism. Such complaints are often supported by suggesting that real professionals would exhibit more concern with getting the job done well than with minimizing the hours of work. For example, it is difficult to reconcile professionalism with the insistence that inservice training occur during the school year at the expense of student learning time, instead of during the summer; or with pushing for the granting of tenure after over-short probationary periods, instead of relating it to the time it takes to make a sound decision[1].

3. The Limits of Negotiation

Since political parameters vary from state to state, from public to private school systems, and from district to district, there is little here about political compromises, although they must be addressed on a local basis before implementation. But there is a special problem about those compromises which have resulted in contracts or conditions of work that incorporate improper, invalid or inadequate pro-

[1] Dismal though the status of teaching may be at the college level, it is at least true that a seven-year probationary period is still generally applied, whereas schools are moving towards one year (a nominal two years often amounts to one if the 'advanced warning' requirements are stringent). Teachers themselves often say that it took them four or five years to develop even a sense of *confidence*, and many say that a sense of *mastery* only came in the seventh to tenth year.

cedures[1]. Since, by definition, such compromises involve the abandonment of full accountability and of full professionalism, they should be eliminated as soon as possible; removing them should be the leading items in the district's negotiation priorities. The aim of schools is to provide the best service to students and community, along with the most effective use of the school's resources, within the constraints of justice and concern for staff. There is no room in that formula for any party to avoid accountability or professional development.

By contrast with improper compromises forced on the system under the threat of strike, the process of serious discussion of any proposed evaluation system with teachers—and/or with their representatives, if they have an organization—is ethically mandatory, and often leads to improvements that benefit all parties. Teachers must be treated as legitimately concerned employees who will be significantly affected by the proposals; in return, they must treat administrators as legitimately concerned to ensure accountability, and to optimize student benefits and staff competence.

4. Administrator Accountability

It is clear that a major theme of the approach described here is accountability, which might be summarized as the view that 'responsibility includes demonstrability'. But, to show responsible use of resources by a school system requires an evaluation of their use by *each* of the various components. Accountability is best seen as a property of whole systems, not of sub-systems. It is difficult to enforce much accountability on one sub-system if you can't tell how much of what happens there is due to deficiencies in some other sub-system which is *not* being checked. In particular, there can be no full accountability of teachers without accountability of administrators. This is partly because teachers' efficiency—and hence their performance when evaluated—is dependent on the administrators' provision of services and functions that are often lacking (for example, in dealing with troublemakers). But it is also ethically objectionable to expect the teachers in a system to commit to a discipline which the administrators avoid, although they need it just as much and the community has the same right to it. (This line of thought has led many school boards, with good reason, to arrange that they themselves be externally evaluated.)

---

[1] Examples of each of these errors, in the order given: the exclusion of *all* consideration of student performance from teacher evaluation procedures; the requirement that evaluation be based on a small number of *pre-announced* classroom visits (or on visits all but one of which are preannounced); the prohibition of *all* review of those teachers who are not applying for promotion or raises. If such conditions are in force—as they are in many districts and in some states—they should be negotiated out as soon as possible. Their presence is a scandal—an attempt to legitimate simple exploitation of the children and the taxpayers. One might as well have contracts with textbook publishers or computer vendors prohibiting the dropping of their products just because they have become seriously out of date. Of course, since few teachers and administrators—or the professors of education who taught them—understand the issues involved in the validity of teacher evaluation, many of these requirements were not recognized as illicit when they were incorporated in the contract, so venal *motives* were usually not involved. The same is true of many peddlers of fake cures for cancer.

5. Superficial vs. Serious Administrator Evaluation

School boards often think that they see enough of the superintendent to be able to assess her or his performance without any *formal* process of evaluation. That's their first mistake, as many have found out later when some of the behind-the-scenes scandals surface[1]. And boards are often persuaded by the superintendent that s/he is closely supervising the other administrators in the same way that the Board is supervising the superintendent. That's a second mistake. Both of those approaches to administrator evaluation are naive, erratic, and quite inappropriately informal for the kind of public-money budget involved. If a board is going to demand more in the way of *teacher* evaluation than the principal's word that it is being done well, they must apply the same logic at the *administrator* level.

Hence this discussion of how to ensure the accountability of teachers in a just and justifiable way is undertaken on the assumption that there is a sound process of administrator evaluation in place or being put in place, not just promised. Such a system is not very difficult to set up, but it is a moral imperative that, simple or difficult, it should be set up no later than one imposes a corresponding system on the teachers[2].

It's quite common for an *unsound* process of administrator evaluation to be in place. One approach, of which administrators and boards seem to be especially proud, is some version of MBO (Management by Objectives). This cannot be taken seriously as a system of evaluation, since it lacks input from those most affected by the administrator's performance, and corrupts what it evaluates by not rewarding the use of 'targets of opportunity', since they are, by definition, not in the annual plan. On the first point, it commits the error which administrators quickly identify in the suggestion that teacher evaluation be done by the teacher's peers (in the sense of fellow-administrators); they are less quick to see the problem when it benefits them[3].

---

[1] One reason that boards get caught on this point is that they are often used to the business environment where what appears to be informal evaluation of administrators is quite common. But there is a key difference: in the business environment, there is a background criterion which provides an objective measure of overwhelming importance—the profit figures. What looks informal is in fact quantitative and very tough. In the world of education, the armed services, and other public services, an analogous backbone must be provided to any overtly informal evaluation system.

[2] Of course, it's quite common for an *unsound* process of administrator evaluation to be in place. One type of which administrators seem to be proud is some version of MBO (Management by Objectives). As a system of evaluation, this cannot be taken seriously, since it lacks input from those most affected by the administrator's performance, and corrupts what it evaluates by not rewarding the use of 'targets of opportunity', since they are, by definition, not in the annual plan. (Some of their other problems are discussed in ***Evaluation Thesaurus***, (Third Edition, Edgepress, 1981)). The hot entry in the administrator evaluation stakes is some version of a simulation test, which has the main merit of being better than the third entry, the 'research-based' approach. Good simulations can provide a not-too-reliable idea of performance under high motivation on some dimensions, for those who do respond well to faked situations where they know they are being watched and judged; but since what one needs is a measure of everyday performance, that's not much. (Interviews suffer from similar defects.) And simulations are a poor test for real courage, loyalty, and tenacity, since everyone knows they are simulations and of limited length.

[3] Some of the other problems with this approach are discussed in ***Evaluation Thesaurus***, (Third Edition, Edgepress, 1981).

Probably the hottest entry in the administrator evaluation stakes at the moment is some version of a simulation test. This has, as far as one can see, only one significant merit, namely that it is somewhat better than the third entry, the 'research-based' approach. Simulations generate not-too-reliable judgements of performance under high motivation on some dimensions, for those who happen to respond well to faked situations where they know they are being watched and judged. But since what one needs is a reliable measure of everyday performance, that's not saying much. (Interviews suffer from similar defects.) Simulations are an especially poor test for real courage, loyalty, and tenacity, since everyone knows they are simulations and of limited length. The same points apply to their use for teacher evaluation. The problems with the 'research-based' approach to administrator evaluation are similar to those discussed below in the case of its use for teacher evaluation.

6. The Connection to Professional Development

While not an imperative to quite the same degree as the imperative to have a serious system for supporting personnel decisions, it is desirable and often politically important for a system of teacher evaluation to have *associated with it* a correlated system aimed at improvement. (We'll call such systems 'enriched' by comparison with 'core' systems.) Recommendations for improvement in teaching typically require more detailed 'diagnostic evaluation'[1] and a different kind of knowledge than personnel decisions[2], just as prescriptions for treating a serious illness typically require more careful examination and testing, and knowledge, than is involved in determining that the patient *is* seriously ill—the latter situation may be obvious at a glance. Professional development can also be facilitated by the provision of resources, especially reference works, cassette recorders and if possible videotaping resources, to assist the individual in self-evaluation and in responding to suggestions and experimentation.

The claim that opens this section is very carefully phrased. It should be sharply distinguished from the oft-made claim that systems of teacher evaluation should be *aimed at* improvement rather than at necessary personnel decisions; or the claim that negative judgements about competence are *invalid* unless backed up by a list of specific recommendations that, if followed, would produce acceptable performance. One might as well argue that one cannot validly conclude that a typist has failed a typing

---

[1] Sometimes called 'formative evaluation'.

[2] The kind of evaluation required for personnel decisions is sometimes called 'summative evaluation'. Thus, a core system always and only provides for summative evaluation, whereas an enriched system provides both formative and summative evaluation *and* training-process recommendations. Sometimes, but only sometimes, the best summative process will spin off *some* useful formative insights. Usually you need more than that, and should do some further formative evaluation. Sometimes, too, the summative report obviously implies that certain training and practice would be desirable. It's rare and accidental if it implies a comprehensive list of what's needed in the way of training and practice. For that you need the further details from a formative evaluation and the special expertise of a trainer skilled in this area.

test unless one can work out a training formula for getting them through it. The task of personnel evaluation is to evaluate the performance of personnel: the task of training or development is related but essentially different—to alter behavior so that it will pass the evaluation.

Providing remediation is highly desirable from the employees' point of view, but it is *a further enterprise* that goes beyond developing a valid system for justifying personnel decisions. And it goes beyond what is provided for most professionals and most industrial workers. Essentially, it is a decision for the school board whether they feel it is appropriate to spend the extra money for a system of detailed diagnosis and remediation rather than dismiss incompetents and replace them. (The evidence suggests that it is more expensive to remediate.) Educational practice has been to go for what is referred to here as an enriched system of teacher evaluation—and administrator evaluation—which takes the extra step. And if a new system of personnel evaluation is being introduced, then fairness *in that process* might well be said to require provision for remediation, since 'the rules have been changed in midstream'. (In reality, it's only that 'the rules' are being applied properly rather than improperly or not at all.) But, once the changeover has occurred, the fact is that a fair system of summative evaluation does not have to be an enriched system.

Notice the crucial difference between the two absolute requirements that do apply to a system of personnel evaluation, and the one that does not; the courts and the unions have frequently confused them.

(i) It is absolutely necessary that there be *solid evidence* for any negative conclusion[1] on which an adverse personnel action is to be based; and this evidence should go beyond the personal judgement of one person, however experienced or well credentialed[2].

(ii) It is essential that negative conclusions *identify the details of the deficiency in performance*, meaning the dimensions along which performance was inadequate and the extent of the inadequacy on each (assuming that there is more than one[3]). In the typing case, for example, someone who fails should be told the extent to which the failure was due to (a) speed deficiency, and (b) accuracy deficiency. Where several dimensions are involved, their relative weighting should also be specified and it should be stated whether there are 'absolute minimum scores' on any of them, that is, scores below which deficiencies cannot be traded off by scoring high on other dimensions.

---

[1] The focus on negative conclusions is simply because these are the job-threatening ones and hence the most serious for all concerned; the logic is just the same when justifying 'award' conclusions.

[2] To a layman's eye, the courts' line of thought suggests that it might be enough for a single person to have *witnessed* several serious transgressions; it isn't enough for the case to rest on the interpretation ('judgement') of a single person, and it may not be enough for it to rest on the judgements of several such people, if the amount of judgement is considerable and/or of debatable validity.

[3] In rare cases, there is a single 'holistic' criterion which has been validated, but can't be fractionated; the best-known case in education is the holistic marking of English compositions. In such cases, there is no obligation to fractionate under appeal, since there are no fractions (scores on separate dimensions) to reveal.

(iii) It is *not* relevant to the validity, demonstrability, or fairness of a personnel evaluation that one identify either the cause of the deficiency or a set of procedures which, if followed, will eliminate it.

Another way to bring out the differences between legitimate and illegitimate demands on the evaluation system is expressed in the italics of the following sentence. "While the system proposed here, and most of the invalid systems currently in place, readily spin-off *directions* along which one needs to move for improvement (that is, the dimensions of the deficiency—along with its magnitude), from which one can easily *infer* in *most* cases how to go about *improving* one's performance, that is very different from providing a *guaranteed* and *comprehensive* remediation procedure, let alone the kind of *support and training* that will ensure success *at the time of the final review.*"[1]

Teaching ability is very situation-specific. It should never be argued that because someone is not very good as a teacher of twelfth-grade history students (or retarded eighth-grade students in a history course, or graduate students in a history seminar) that they are 'not a good teacher', or 'not a good history teacher'. There are plenty of counter-examples to that kind of overgeneralization. Conversely, it should never be assumed that any particular teacher, even one that has an outstanding track record elsewhere or in earlier years, can be brought up to the level of acceptable standards in every possible teaching job in which they have been placed or even for which they have been credentialed. Some teachers are simply not suitable *and never will be suitable* for a particular job within their field of certification, although excellent in others. (There is a close analogy with the tasks of research and acting.)

Hence it is absurd to suggest that an evaluator must always be able to indicate to someone being evaluated just how that person could become a good instead of a poor teacher *in their present situation*; Doing that is sometimes impossible. The appropriate requirement is simply that the evaluator be able to indicate the dimensions and magnitude of the deficiency. Any more than that must come from someone with remediation knowledge and skills, and can only be expected to maximize the chances of success, not ensure it.

7. Segregating Summative Evaluation from Development Support

The Ideal In an enriched system—one that provides remediation support and not just summative evaluation—the results of the further (formative) evaluation that ensues when it becomes clear that remediation is necessary, *and* the recommendations for learning/training/practice, *and* any observations on progress should be fully segregated from the summative evaluation. If this is not done, it will

---

[1] In light of what is said in the sentence in quotes, it will be observed that the cost of a 'enriched' evaluation system lies—as it happens—mainly in the cost of the teacher-trainer rather than in the need for a separate system of more detailed diagnosis. But only 'mainly'; there will be plenty of cases where the further formative evaluation will be important before appropriate training/practice can be recommended.

obviously not be reasonable to expect much willingness on the part of teacher-clients to go to formative advisors about what they perceive as their weaknesses. One might as well expect that clients would seek advice from attorneys if the latter are doubling-up as judges on the same case; and it's not just the guilty ones that would stay away. Furthermore, if teachers *do* get help from the person who will judge them, we have a classic case of 'teaching to the test' combined with the lack of objectivity implicit in the idea of an author reviewing his or her own work. From the community's point of view, it is a very poor way to ensure objective evaluation. Yet it has been the norm for centuries. It is one relic of the little red schoolhouse that we should try to transcend.

While the recommended role-segregation of the summative evaluator from the formative evaluator and helper has not been the tradition, the change to separation of the roles does not have to involve a large expense in larger schools, since it can be set up so that it mainly involves a different way of slicing the administrative workload. The principal, in most systems, has the responsibility for making and defending the personnel decisions (even if they have to be approved by the superintendent and sometimes the board), and is hence the natural choice for the summative evaluator. In some school systems, an assistant or deputy principal can then be put in charge of the support system.

There are advantages to using an outside consultant in that role—and it may not be much more costly—because the likelihood of an assistant principal letting information slip, or having to refuse a direct though improper request from a superior for confidential information, is avoided. Credibility is still further advanced and security breaches not made much more likely by using 'mentor teachers' in this role—outstanding teachers who are still on the payroll as teachers, in the same or another school. The mentor position is an appointment that requires serious evaluation and which pays off in increased salary and prestige (details are covered in a later section). In all these cases—deputy principal, consultant, and mentor—performance of the staff development part of their job is rated almost entirely by the teachers who are served, the role as developer should only be for one to three years although renewable, and violation of confidence is grounds for dismissal. In addition, any allegation of such a violation that is not completely frivolous should result in a hearing.

The Reality Isn't all this too squeamish? After all, the system of having supervisors make (summative) recommendations on their supervisees, as well as assisting them to improve, is widespread in the helping professions. And teachers all have to serve in the summative as well as the formative mode for their own students; as do parents. Isn't it part of the human condition to have to cope with multiple roles?

Of course it is; but it leads to poor performance in many situations, and should be avoided if possible. The key question is how much the school system is interested in improving current staff performance. If it is seriously interested, it has to deal with the fact that many teachers won't go to a principal for help—because it is likely to disadvantage them—and don't like to go to their peers, or can't get help

from them. (Alternatively, they go to the principal *in order to* influence her or him by getting them involved in, and hence co-authoring, the solution to their problem.)

The counselor/judge role conflict has spawned many remedies and no panaceas. In the academic area, we have long used external exams as a way to free teachers to be allies with students in the common cause of success in the exams. Parents have often divided the disciplinary and the supportive role between them, or between them and the school. We recognize the problem everywhere.

In the professions, we often allow (or, for practical reasons, have to allow) a period of apprenticeship—for the intern at the hospital, the practice or probationary teacher, for the articled clerk—when close supervision is essential and professional or personal rivalry is scarcely relevant. Even during that period we try to provide at least some independent assessment. And we usually move towards a different approach as autonomy develops. We look for objective measures of performance—the post-mortem and the organ committee for the doctor, the publications list for the researcher, the ratio of successes in court for the attorney, juried awards to the writer and architect. But in the schools, we have failed to do this. The usual concession to the experienced teacher is make the classroom visits perfunctory, or to omit them entirely. This is to deprive the teacher of what should be valued and valuable information[1]. It also deprives the community of the improved teaching that should result from regular evaluation, and of insurance against keeping teachers on the staff when they are no longer doing an acceptable job.

Practical Compromises We need to bring the reality closer to the ideal, if we wish to improve teaching. If resources will not permit a division of labor, or the use of consultants or mentors, then we must simply do the best we can to help the principal separate the roles internally. Some steps in that direction are: the use of standardized forms on which only legitimate dimensions are rated (the basis for those are provided in this article); staff meetings a few times a year at which the principal can express views without putting pressure on an individual to do what the principal says instead of doing what the job requires; providing strict training for all principals in the difference between what they can legitimately 'observe' in a classroom, and what they *must disregard* (this distinction is explained later); providing good print or courseware resources for teachers, and some subsidy if they will attend inservice seminars, so that they can get answers to their questions about teaching and thus avoid loss of face to peers or loss of rating from the principal; providing sound and video recorders so that teachers can self-evaluate; rewarding teachers explicitly for helping peers; and so on.

---

[1] The idea that teachers will benefit from regular summative evaluation is often thought to be naive or deceptive. Far from it. Although one hears the complaint often enough in the US, it was striking how many teachers in Western Australia volunteered submissions to the Beazley Committee in 1983 lamenting the fact that in twenty years no-one had ever bothered to visit their classroom, or talk about their problems and questions about teaching. What are we telling such teachers? We may think that we are telling them that we trust them; ' since everyone knows that some of the long-term teachers are complete disasters, what we are really telling them is that we don't care enough about the education of students or the professional development of teachers to check on the process.

8. Teacher Competency Testing vs. Teacher Evaluation

The most important need in the schools is for a comprehensive approach to teacher evaluation. Nevertheless, one should not imply that all piecemeal approaches are without value. Properly used, they can avoid some major disasters for student and schools, and they can be put in place fast, fairly, and cheaply. But one must understand the logic of such approaches very clearly, or the fairness will be absent. Fast and cheap is no substitute for fair and full; but it is sometimes better than nothing. Moreover, while the proper use of these partial approaches is better than nothing, and certainly better than using one of the invalid attempts at overall evaluation, it is only appropriate as a first step.

The areas in which this approach has been most popular relate to tests of minimum competence in literacy, numeracy, knowledge of subject-matter in the teacher's area(s), and pedagogy. Of these, the last one raises some serious problems. A major effort was made a number of years ago to develop a complete set of pedagogical competencies, and to base teacher education curricula on them (this was called "competency-based teacher education"[1]). It turned out to be somewhat ahead of the research of its time and for the reasons given here not a usable basis for overall teacher appraisal. But there are areas of pedagogy where little debate about the proper use of research results is involved. Checking on test-construction/interpretation skills and minimum competence in the use of simple audio-visual devices—and perhaps even of computers—might be candidates. This is not because every teacher should *use* all of these skills, but because—it might be argued—everyone should know *how to use* them so as to be able to take advantage of their special benefits when and if an opportunity arises.

The current development of more sophisticated competency tests by the Carnegie Project at Stanford (directed by Lee Shulman), can be expected to increase our repertoire here, in both quality and quantity. In the end, however, testing of that kind can only provide a partial approach to teacher evaluation, as to administrator evaluation. One must have a great deal more, such as input from students and parents. The Carnegie team has shown some interest in a more comprehensive approach, and may go on to give us all some guidance in that area, too.

However, the first point to make about the logic of such tests, measures or observations (we use the term 'test' for convenience) is that they can only be used to support unfavorable actions, not favorable ones. There is nothing unfair about this, though that comment is widely made by teachers. Good secretaries are more than good typists, but no-one ever suggested that makes it unfair to use a typing test as a screening test. The same logic applies at mid-career: there is much more to being a good surgeon

---

[1] Some comments on the approach will be found in "Evaluating Program Effectiveness or, If the Program is Competency-Based, How Come the Evaluation is Costing So Much?", ERIC Document No.SP008 235 (ED 093866), in ***Research in Education***, November, 1974.

than good diagnosis, but no-one ever suggested we should keep surgeons on staff who keep removing healthy organs.

This asymmetry about the kinds of evaluative conclusions that can be obtained from isolated competency testing is one major reason why their use should only be a short-run policy, and one reason why the main focus of this paper is on complete systems of evaluation.

The conditions that must be met on such tests include the following. The test must be: (i) a valid way to identify the presence of (ii) a minimum level of competence that is essential for competent teaching. It must also: (iii) be applied uniformly to all those for whom it is relevant; and (iv) it can only be selected from a range of such tests on grounds that pre-empt deliberate discrimination, such as feasibility or demonstrated deficiency. (De facto discrimination is not an argument against validity, here or elsewhere, as long as condition (iv) is met).

By failing to test all the other dimensions of good teaching, the competency approach fails as a procedure for the identification of good teachers. But if it does test one minimum level of competency on one dimension, it can screen out people who should not be in the profession at all. Their deficiency in this dimension cannot be made up by performance on other dimensions (as condition (ii) makes clear), so the fact that performance on the other dimensions is not tested is irrelevant to the validity of using the competency test for dismissal or remediation decisions (as long as condition (iv) holds).

It must be realized that the adequacy as opposed to the validity of competency testing needs separate consideration. A test of teacher literacy that is perfectly valid in principle may be used in such a way as to provide no insurance against incompetence. If, for example, the cutting score for passing is set too low (Texas), or if unlimited retesting is allowed despite the existence of only a small pool of items (Florida), then incompetent teachers will get through. The politics of testing is such that it always raises a storm of controversy, usually because of differential racial impact, from which the politicians are likely to retreat without having really seen the job through. The only counter to this is for parents and employers to keep the pressure on, and to do that, they need to understand the kind of issue discussed here. It's part of enlightened citizenship and parenting training today.

9. Pro-Teacher vs. Pro-Administrator Evaluation Systems

Those of us who have proposed systems for teacher evaluation are often asked if our own system is 'pro-teacher' or 'pro-administrator' (the latter being taken to mean anti-teacher). And no doubt we all respond by saying that it's neutral. In the present case the argument for balance might include the following points. On the one hand, the approach provides a very strong defence of the teacher against unjust methods of evaluation, which includes virtually all current approaches. Furthermore, this system insists on the complete right of teachers as professionals to select the approach to teaching that they find best-suited to their own character, talents, students and subject-matter; and, once tenured,

on their right to disregard, without penalty, suggestions from their superiors as to what they should do when the need for improvement has been demonstrated. And it argues strongly for a support system which is segregated as well as strong, the first requirement being a way to eliminate a very important source of injustice.

On the other hand, the approach taken here provides a very strong defence of the student, parent, and taxpayer against exploitation by the lazy or incompetent teacher. Amongst other things, this approach completely rejects the idea of resting teacher evaluation mainly on peer assessments, or on the assumption that someone who was once competent is forever competent, or on a request for promotion, or on the ability to specify remediation procedures, or on the overall, undetailed, judgement of a principal who may be weak or a crony. While it defends the right of the teacher to disregard the principal's advice, it does nothing to soften the need to meet the standards, however the preparation for that is undertaken.

So this approach is not intended to, nor does it, cast anyone to the wolves or leave anyone else with no protection against wolves—or mules.

II: THE NEED FOR A NEW SYSTEM

10. Introduction

We need a new system because all the old ones are not only inefficient but invalid, and hence unjust (since an invalid system will award benefits/penalties to people who do not deserve them). Even the systems built into the latest handbooks for teacher evaluation, issued by a number of states in the US that have laid out a great deal of money to ensure that their new approach is based on the latest research—Georgia is one example—are illegitimate. If the arguments presented here are correct, we will have to expect that most existing systems will be thrown out by the courts. Colossal retroactive damages are a possibility in many jurisdictions.

The reasons for rejecting existing systems will be reviewed next, prior to outlining the one that is here suggested as a viable alternative.

Initially, the focus of the discussion is on the kind of teacher evaluation that is required as a sound basis for making personnel decisions (summative evaluation). This is just as essential as a starting point for planned professional development as it is for personnel decisions. We have already indicated that there is sometimes a need to go further—into formative evaluation—in order to support development, and some suggestions are provided as to how that can be done. However, summative evaluation may and often does give us all we need for the development process—in which case it serves as formative as well as summative evaluation.

It is also stressed here that it is risky to start a professional development process unless negative results from a reliable summative evaluation are at hand. This is not just on the principle "if it works, don't fix it" but because improvements in one dimension of teacher performance can produce deterioration in other and more important areas. So the natural interest of all true professionals in improvement, no matter whether they are already good, must be pursued cautiously, and not in the all-too-common spirit of trying to incorporate and raise enthusiasm about every new fashion in teaching.

The first target in the following treatment is the latest and apparently most respectable approach, 'research-based teacher evaluation' (RBTE). It is best to begin with this one since its advocates are already convinced of the inadequacy of the other approaches, and would not find much interest in starting with a survey of their problems. But if the new approach does not work, then the other approaches need to be re-examined carefully. We have a great deal of experience with them, and if one of them can be salvaged, it would be preferable to use it instead of switching to a new one.

11. 'Research-Based' Teacher Evaluation

We begin by proving that in personnel evaluation one cannot use any of the research on teaching that has allegedly shown certain teaching styles[1] to be more effective than others. It had been supposed, naturally enough, that this research meant we could use classroom visits to see which teachers exemplify these 'winning styles' best, and use the results of those observations to select (or promote or retain) better teachers. But it is argued here that we cannot use any such observations in any such way. The essential problem is not that the generalizations aren't true; although there are some grounds for concern about their truth[2], we put those concerns aside for the purposes of this discussion, because they are not the essential problem with RBTE. To understand the real problem, we can start with a case where we have learnt to appreciate the impropriety of using empirically sound generalizations in personnel decisions.

Considerations of Justice Let us suppose that it is true that women tend, statistically speaking, to make better primary school teachers than men. That fact would still not justify one in using gender to select a female over a male applicant for a post as a primary school teacher. We need to understand the basic reasons against using gender in such as situation. It is not that it happens to be illegal in many

---

[1] The concept of style as it is used here can best be defined by example. The styles of teaching that RBTE usually favors include: the use of advance organizers (lesson objectives put on the board, handed out, or mentioned), asking questions of students, encouraging students to ask questions, frequently providing positive reinforcement, maintaining eye contact, maintaining high 'time on task'. The failure to employ or exhibit any or all of these is also a style. On the other hand, treating students justly, providing them with feedback on their academic progress, explaining material in language that they can understand, and giving them clear directions about assignments, are not matters of style. They are duties.

[2] The most serious complaint is that they involve illicit generalization from small studies to other regions, student types, and subject matters; but there is also a problem about the definition of good teaching that many of them involve, since it frequently omits long-term retention measures and performance on out of class duties.

jurisdictions as a violation of affirmative action legislation. It is not that it would be politically unpopular with men. The first fundamental reason is that selection for jobs by gender is, in all usual cases, a simple violation of natural justice. It involves 'guilt by association', that is, it involves penalizing someone—in this case, any male candidate— because of the average performance of a group to which he happens to belong, instead of judging him on his own individual merit. Even if women are fifty times more likely than men—on the average—to make good primary teachers, you must, ethically speaking (as well as legally), base your decision solely on the legal credentials and track record[1] of the individual male and female candidates.

We all have some feeling for that point when it's applied to gender or racial or religious discrimination, but it applies equally to *all personnel decisions made on the basis of any generalization about a group to which the applicant belongs*, including the group which teaches in a certain way. Hence, using any use of evidence about the teaching *style* as opposed to the teaching *achievements* of this individual, is a violation of natural justice.

Apart from natural justice, there are both economic (efficiency) and scientific reasons for not using style data.

Considerations of Efficiency Of course, if the correlations between gender and merit are extremely high and if it was very expensive to get or interpret track record data on the individuals, it might be more *costly* to follow the path of justice (it all depends on how you estimate the costs of mistakes vs. the benefits of successes[2]). Nevertheless, we would have to do it for the same reason that we pay the bill for public defenders—because justice outranks economy. But in fact all the correlations we are talking about are modest in size, they cannot be combined in order to obtain higher correlations (because we have no reason to think them statistically independent of each other, the precondition for amalgamation), and for the most part we already have or can easily get the track record data. So the

---

[1] We use the term 'track record' to refer to data about prior achievements by the candidate in the same or similar jobs. And of course 'achievement' refers to the discharge of duties, not to the display of a particular style in discharging them (except where the style is part of the duties, as with divers and dancers).

[2] The problem is that the costs are different if one calculates them from the point of view of the administrator by contrast with the point of view of the applicants, the children, or the taxpayers. For the administrator, it is very stressful to have to cope with the complaints about a bad teacher, and possibly with the enormous effort involved in trying to dismiss with cause. So the administrator tends to clutch at straws that might reduce what s/he sees as bad choices; technically, this is the effort to reduce 'false positives'. By contrast, passing over a good teacher (because s/he happens to have a style that is not typical of good teachers) doesn't show up as a cost to the administrator—unless *no* candidates pass the Approved Style Test. But in fact missing a good teacher is just as much a bad choice, just as much an error; a false negative. It's just that the cost is not paid by the administrator; it will be paid by the candidate who doesn't get the job, and by the students, who will not get the best teacher from amongst the applicants.

One key point for the administrator to remember is that as long as other administrators are using the Approved Style Test (AST), a strong competitive advantage is gained by ignoring it, since the pool which is passed over will contain some and possibly all of the best candidates, if you are prepared to judge them on their own merit. Remember that those not prejudiced against the award of post-doctoral fellowships to women finished up with better scholars; and the same was true with stock analysts, news directors, and airline pilots. Of course, it has to be shown that using the AST does not give an advantage; and that's what this section is about.

special cases in which you might try to justify using the generalizations to save money or time do not apply. The simple truth is that track record data is a better predictor than any other that you can get hold of.

But surely we can at least use *both* kinds of data—simply combine the track record data with the style data? No, because: (i) that's still unjust, since it makes part of the judgement depend on data that isn't about this individual candidate; (ii) these two types of data are not statistically independent and hence can't be combined; (iii) possessing the track record data automatically makes the other generalization irrelevant, because the overall generalization refers to *random* samples from a certain defined population, and what you know about the candidates shows they are *not* random samples from that population.

Of course, *if* we had to make personnel decisions in one second, based on *one* fact about the applicants (gender or skin color), *other than* track record, and *if* we exclude social and ethical considerations, we'd do well to jump with the generalization. "Women and children first" makes sense when the Titanic is sinking and there's no time to look for a better basis for selection. It would take more time than we have to identify any researchers on board who are close to a cure for cancer. In personnel selection, we have plenty of time to do better than the generalization. Using the generalization is akin to buying a used Honda Civic without driving it, on the grounds that the road tests identify it as the best car in its class. That's just bad practice, in efficiency terms. And it's bad scientific method; you have failed to gather some evidence that you could easily gather, evidence that will greatly increase the chance of making a successful prediction.

Considerations of Scientific Method The key point here is that scientific method covers data gathering not just data processing. You are just as guilty of poor scientific method if you don't get evidence that's obviously relevant as if you draw the wrong inference from the evidence you do have. If you want to select the best candidate, you must get hold of the best evidence that is available at reasonable costs in time or money, that is, costs that will be more than reimbursed by the gains in quality or savings in money that will result from getting, retaining, or promoting the best candidate.

If a doctor examining a male patient decided not to call for a biopsy on a chest lump because it's statistically unlikely for males to get breast cancer, would be guilty of scientific error. Making personnel decisions on the basis of the style generalizations is exactly the same.

The First Underlying Fallacy Our thinking about such matters has been careless because of two attractive fallacies of oversimplification. The first fallacy is the supposition that the error underlying discrimination is an error of fact. We have become accustomed to reject discrimination against women candidates for school administration positions on the simple grounds that there's no evidence they *are* inferior. That appears to be the case, but in a sense it's the easy way out; it's not the essential point. For it doesn't matter whether 'they' are inferior; that would still be no justification for the discrimination. It's perfectly clear that *some* women are better administrators than *some* men, just as the

reverse is perfectly clear, and that alone means you have to look at the individual cases. That's what justice requires. But science requires the same, though it's less easy to see this point.

Even if *no* women were better administrators than *any* men—that is, if the correlation between masculinity and merit is 1.00—you still can't use gender as an indicator, because the evidence on which any such generalization is based necessarily refers to past cases and you can't assume that it will be true of the next case you run into. The cultural and media environments change, and consciousness changes, and then laws and training programs change, women change, men change, jobs change. Whatever generalizations about the relative merit of men and women for certain jobs happen to be true at any moment are only historical summaries, not a usable basis for future decisions. In many recent cases, people have become very good at jobs that members of their sex had never previously attempted—in their own culture. We will miss all such cases if we predict on the basis of the historical generalization, and those misses are not only unjust but wasteful.

Given the size of social and personal changes across a period of a few years, and the size of the human interest in falsifying generalizations about humans or sub-groups of humans, and the fact that doing something different is essential for new fashions, it is essentially certain that generalizations will always be falsified, so it is essentially certain that those who use them will make mistakes and hence commit injustices and waste resources.

In the real situation, where the original discoveries are weak statistical tendencies, which could even reverse their sign[1] in the course of a few years, there is a high chance of error and injustice from using them. The essential flaw in discrimination is thus not an erroneous assumption of fact; it involves a completely fallacious method.

A particularly nasty feature of the use of style, sex or race for personnel selection is its 'self-perpetuating' aspect. If we continue to use these factors as negative indicators, we make it impossible to discover that they are unsound indicators because we never allow potential counter-examples the opportunity to show that they are counter-examples. This is an error of method, in terms of social procedures if not scientific method. Correct procedure—if there were nothing else wrong with the use of the indicators—would involve occasional experimentation to test the continuing validity of the generalization.

Summary Discrimination is not wrong because of some irritating politically-motivated intervention by the government, or because it's wrong to disobey the law, or politically risky to alienate women or minorities, or 'time that we compensated for past injustice by biasing things in the other direction for a while'. It is wrong because it is, in the first place, *intrinsically unjust*. It is wrong because it is, in the second place, *cost-ineffective*. It is wrong because, in the third place, it usually involves a scien-

---

[1] That is, go from being positive to negative correlations.

tifically unsound procedure (*failing to gather the evidence* that will make possible a more reliable decision). It is wrong *methodologically* because it involves an approach which protects its assumptions from disproof by the facts. And, fifthly and almost incidentally, it is sometimes wrong because it *assumes a falsehood* as fact. The point is that even if the fifth consideration were absent, the approach is completely illicit.

Hence it is impossible to base teacher evaluation on any evidence about the teacher's 'teaching style'. However sound the research may be that shows a certain style to be, on balance, superior to the style of other candidates, it cannot be used as an indicator of merit any more than gender can be used an indicator of merit, even if each is in fact a *statistically valid* indicator. Style, like sex and religion, is an illicit indicator for personnel decisions because it is *only a statistical* indicator of merit.

The Second Fallacy Many of us feel that we should never be made to throw out statistically valid indicators. We suppose that statistically validated indicators always tell us at least 'part of the truth', and one should never have to throw out part of the truth. From this there arises some of the resentment one often finds about civil rights or affirmative action legislation which denies the right to use gender or race as discriminators, whether or not they do actually discriminate as indicators of merit: "Why shouldn't we be allowed to use the facts?" But statistical generalizations do not tell us part of the truth; they tell us *all that is known about some variables* in a situation where *all that we know is the value of some other variables*.

In the situation in which all we know about two candidates is their gender, then the whole of the truth about their teaching ability in primary school is that the women are more likely to be good at it than the men. If all that we know about two candidates is their race, then the whole of the truth about their criminal record is that the blacks are more likely to be criminals than the whites. The catch with this kind of knowledge is that it is totally conditional. It only exists if the conditions are met. The moment you know something else about the candidates, it is in jeopardy; and if what you know is in any way relevant, it's gone. Thus, if you know that all the candidates have college degrees, you *no longer know that the blacks are more likely to have a criminal record*. It isn't still there as an extra fact up your sleeve; it's totally irrelevant to this group, because college education markedly affects incidence of criminality. With teachers, the research on winning styles tells you that if *all you know* about two teachers is that one exhibits a winning style and the other does not, the one with the winning style is more likely to be successful. The dilemma is this: if you also know something about their track record, you've violated the condition of ignorance about any other relevant data, so the style generalization no longer applies: and if you don't know anything about their track records, you have violated the condition of conscientious performance of your own duties. In short, the use of style data only justifies one kind of personnel decision—the dismissal of the person that uses it.

The Lazy Use of Style Data The problem in personnel evaluation is that we are using secondary indicators such as classroom style *instead of* direct evidence on the performance of duties, and that is

totally inappropriate. You can't even use them *as well as* direct evidence, in personnel evaluation any more than in medicine. You could only use them when the better evidence is not available, and that, by definition, excludes personnel decision-making because better evidence is always available there. ("Available" doesn't mean that it happens to be on your desk, ready for use. It means that you, or one of the others involved in making the personnel decisions, can get it if you try.) The process of personnel evaluation is a serious one, not the sort of activity where one can say that a few hours of work gathering direct evidence on the competence of a particular candidate is impossible, "so we have to fall back on secondary indicators". Someone's career, and the future welfare of a large number of students, is at stake.

Perhaps the nearest that an administrator comes to an excuse for using style data is in selecting a teacher for a first position. One often hears it said that there is no track record in such cases. But there is, and a great deal of it. Even on classroom performance, there is track record data, if you can extract it from the report or the person of the supervising teachers at the candidates' practicums. If that were not possible, then you could still and would have to arrange a trial class (not an interview, which is no substitute). And there is plenty of other track record data, bearing on the competence in the subject matter and conscientiousness in the discharge of duties, etc. etc. Look at the list of duties given later, and many more emerge.

Style Data as a Contaminant The preceding warnings are not just about guarding against a small source of error in a procedure which is mostly acceptable. There is also the problem of contamination; there is a risk of invalidating the whole of an approach that involves even one appeal to a statistical indicator. The reasoning is the same as that which invalidates a personnel interview at which candidates are asked a question about, say, their private lives. Even if 99% of the information acquired is licit, the response to the one item that is illicit may in fact have been very influential in the decision. This can be so even if no-one present thought that it was influential, and even if its official weighting on the personnel forms was slight or zero. The involvement of illicit indicators 'contaminates the process' because you can't prove they were ignored at the unconscious level. Justice is portrayed as blindfold because the only sure way to avoid bias is not to know the facts that bias.

Summary The preceding arguments immediately invalidate almost all current, especially recent, approaches to teacher evaluation. Characteristically these so-called 'research-based' approaches are just as improper as sexist or racist approaches. While the motives for using them are less unethical they are still morally culpable. We must therefore find an alternative to what appeared to be the most scientific approach.

12. Measurement-Based Teacher Evaluation

The most obvious alternative, one which also appears to offer some of the objectivity of science is to replace *indicators* of success with the 'real thing'—direct measures of learning by the student. There are special cases where this can lead to a usable result, and the courts have shown that they will accept such approaches, but they are relatively rare. The fundamental problem with the approach is that the measures do not tell us about teacher merit as they stand; they simply provide raw data about something that is happening while the teacher is teaching. To get to an evaluative conclusion one must (i) *establish causality*, and (ii) have some *validated standards*, to apply to the data. The standards must be supported by proof that *this* much achievement by these students (in *this* school, in *this* subject, with *this* much background, and *this* amount of parental support,...) represents a *very good* achievement by the teacher—or a very poor one, or somewhere in between.

Even if we have comparative data about how much the students of different teachers in the same school have learnt, based on common tests of a common curriculum, tests that are designed, administered and scored independently of all the teachers (or at least evaluated by someone with the power to modify them), and even if the classes are matched for pre-test ability and intelligence, you can't tell whether any of the teachers are competent or brilliant, you can only tell how they stand on *relative* competence. But comparative conclusions really won't justify most personnel decisions. For example, they won't justify decision about retention and tenure; the worst of a group of teachers may be good, the best may be poor. Strictly speaking, you shouldn't fire someone for relative incompetence, nor should you give them tenure or promotion for relative competence[1]. Thus, while we should certainly try to obtain evidence of comparative learning gains as *part* of the relevant evidence—it provides us with a reality baseline that becomes increasingly important as it becomes more substantial—we have to do more than that.

Apart from the fact that the best conclusions you can get from comparative learning gain scores aren't the ones you really need, it is extremely difficult to meet the conditions mentioned, if we are to use this kind of measurement approach. And there are other, fundamental, worries about the measurement alternative. It is, for example, a major problem that, in its usual form, it doesn't pay much attention to the content, only to the learning gains (it is quantitative but not qualitative, where we need both). Substantial aspects of the content or its interpretation are usually under the control of the teacher, and should surely be evaluated; after all, teaching rests on the value of the content for almost its entire justification (the rest being socialization, gate-keeping, and baby-sitting etc.). And then there's the matter of improper process—injustice in the classroom and so on. There's nothing about that in the learning gain scores. We need to look elsewhere, though we shouldn't throw away all uses of learning gain scores.

---

[1] Nevertheless, a US Circuit Court of Appeal has accepted the plausible argument that if two competent but not extraordinary teachers can get their students to a certain standard, a third teacher whose students come nowhere near that standard is *prima facie* incompetent; that is, in the absence of specific evidence to the contrary, the conclusion stands.

One reason we can't ignore them completely is that we can use them, rather than teacher estimates of them, as an indication of how much a particular category of students *could* learn. If we don't know how much students could have learnt, we can hardly complain or exclaim about how much they did learn. So the extent to which an individual teacher has achieved the potentiality of the students, which is surely a measure of good teaching, can—it appears—partly be inferred from the actuality of their achievements under several teachers. Thus we certainly can learn something useful from learning gains, just as we can learn something useful from the study of a doctor's patients when we are evaluating him or her for meeting professional standards. It's just that we can't learn everything we need from outcome data—and we often can't get them. So this approach does not represent a general solution to the practical problem of finding a teacher evaluation system.

13. The Judgement-Based Approach

Many have felt that we should put the task of teacher evaluation into the hands of experienced teachers—or ex-teachers such as school principals. Let them directly inspect the classroom. Surely the classroom is the best source of data on how well teachers teach, and surely these people are the best judges of good teaching? Well, of course they *might* be good judges, but how would we ever know? For that matter, how would they know? Even if they did know, they might also be over-kindly in their judgements of other teachers; or for that matter overly harsh; or overly affected by personal appearance or considerations of style; or the standards they use might vary enormously from subject to subject, school to school, year to year, as the individual or the committee membership varies. How would we know that? *If you use judges, you have to validate them*—or face the scepticism of the kind of judges you run into in court. Courts have been more forgiving about the judgemental approach than they should have been, probably because it is the traditional approach and better alternatives—and the details of the failings of this approach—have not been presented to the court.

One reason for being very concerned about the use of teachers or principals as evaluators is that they have in the past developed their own way of coping with classes and may find it difficult to *really* (as opposed to verbally) accept the idea that many alternative ways are, on the evidence, every bit as good. Pathfinders for the westward wagon trains, used to their own way through the mountains, are not especially qualified to tell if another route will go through, even if we take them along on it. In the end, we'll need to have some input from experienced teachers; but we can't build everything on that foundation.

Other reasons for concern about any approach based heavily on classroom visits are that the usual number of visits is far too small to provide a statistically adequate sample of what everyone knows to be a time-dependent process; the sample is in any event not a random one because the visitor's presence introduces changes of unknown magnitude; there is the possibility of social/personal bias at work; and the visiting evaluator's subject matter expertise is likely to be limited. Most of these con-

siderations are enough by themselves to rule out the use of classroom visits as a significant. let alone as the main, basis for teacher evaluation.

The track record of peer evaluation must also be mentioned. It has turned out to be extraordinarily difficult to get peers to turn in negative evaluations even where they are unquestionably appropriate. This difficulty has been encountered outside education, for example in the armed forces. Only in the absolute front-rank tertiary institutions does this system work well, and then usually only because it is focussed on research performance rather than teaching[1]. The problem has been called the problem of 'secret contract bias'; everyone who is called on to make judgements is very conscious of the fact that they will be on the receiving end of these judgements on another occasion. The secret contract—not entirely implicit if a union is involved—is that if they go easy on the victims this time around, it is understood that the favor will be returned on the later occasion.

However, there is a group of judges who are ready, willing, and able to assist in the process of teacher evaluation. These are the students themselves, who are best placed of anyone to tell us some important things—for example, how well the subject-matter was made comprehensible to them[2]. Given a modest amount of proper training about the evaluation of teachers, something of considerable value to them for other reasons, students from the middle primary years upward appear be able to do this quite well.

Overall, the judgemental approach is a disappointment when compared to reasonable standards of evidence for personnel decisions.

14. Other Previous Approaches

Some other approaches that have received occasional support should be mentioned in summary, although we have already said something about them.

Tests and Simulations At its best, a useful partial approach which can serve in the both summative and formative roles. But it can never provide a complete answer, because of its unrealiity, to which different subjects (appear to) respond in very different ways.

Management by Objectives (MBO) This approach is flawed by, for example, the lack of input from affected parties who are in a good position to judge certain aspects of performance (in this case, stu-

---

[1] Centra reports that even Carnegie-Mellon University couldn't make it work for the evaluation of teaching ("Colleagues as Raters of Classroom Instruction", ***Journal of Higher Education***, 1975, pp. 327-337).

[2] One can't validate the use of student ratings on the (true) grounds that the ratings correlate with learning gains. That just gives us one more statistical indicator, as invalid as the rest. There are half a dozen other ways to validate them, however, centring around the fact that they are direct observers of their own comprehension or lack of it. See "The Validity of Student Ratings" in ***Instructional Evaluation***, no. 1, 1988.

dents and peers), by the downgrading of targets of opportunity, by its vulnerab to good talkers (the 'car salesman problem'), and by the absence of data on the performance of many critical duties.

III: A VALID ALTERNATIVE

15. The Duties-Based Approach

To find another answer, we need to go back to fundamentals. We must ask two questions. What is a teacher hired to do; and how can we decide whether it has been done adequately or excellently? The teacher's **primary duty** is to teach students worthwhile knowledge[1] (cognitive, affective, or psychomotor, depending on the teacher's responsibilities) to the extent of the students' abilities[2]. Of course, this is normally understood to imply satisfactory performance in the classroom, but it also entails an obligation to perform a number of **constitutive tasks** that are not part of the central classroom process, such as correcting homework, keeping up to date with the subject-matter, pedagogy and student needs. But there are in addition many **secondary duties**, not required in order to perform the primary duties, such as talking to parents, supervising corridors or lunch-rooms, doing committee work, submitting information on the students' performance to the school administration, referring students to appropriate counsellors, etc. These vary to some extent from site to site, but there are usually a good many of them that all teachers at a site understand to be part of the job. They are no more dispensable for being secondary; they are simply ancillary.

To put it simply, the duties-based approach identifies all of these duties, uses multiple measures to get a best estimate of the extent to which they have been done well, and synthesizes the results. It never uses style indicators, it never relies on a judge for anything that that judge can't be shown to be able validly to judge, and it never confuses comparative merit with criterion-referenced merit. Its validity derives from one source, the obligation of the employee to discharge the duties of the job to the extent

---

[1] Sometimes the content of the curriculum that a teacher is required to teach is not worthwhile or not deemed so by the teacher; an example would be teaching creationism as a reasonable alternative to evolutionary theory as required in some Southern states. The teaching is then an undesirable duty; it still continues to be a duty as long as what is being taught is not so evil or damaging and so rigidly required that the teacher should abandon the job rather than continue to teach this material or ignore the requirement (anti-Semitism in Nazi Germany, for example).

[2] This is (sometimes) ideally done by inspiring them to enjoy the process of learning this material and of learning in general; but that is not always possible given the constraints of entering attitudes, time and resources, and hence cannot be part of the primary duty. (But it might be argued that it is a duty to try this, where possible, rather than assuming it is impossible.) Similarly, while it may be ideal to encourage students to learn to manage their own learning process, and hence that may be something which we should spend some time teaching them, it is not a substitute for the primary duty, which is to teach them the substantial content of the curriculum.

that is reasonably possible with the resources available[1]. This source is unimpeachable on logical, legal and ethical grounds.

Duties-based evaluation can be done exhaustively, in which case it is extremely time-consuming and intrusive; or it can be done pragmatically, in which case its costs are manageable and the results still valid. Nevertheless, it does take more time than the present superficial approach which is legally and scientifically unsound. We should resign ourselves to the fact that we are going to have to put somewhat more resources into staff evaluation and development. Personnel evaluation and development is, after all, the most crucial part of quality control and staff development in the school, and has to be the most important substantial task of the school administrator.

16. The Basic Dimensions of Teacher Merit

The merit of a teacher can, it is suggested here, be exhaustively categorized in terms of just four macro-dimensions, that is, large categories:

A. The **Quality** of the content of the teacher's materials and of the students' learning[2].

B. The **Quantity** of their learning[3].

C. The **Professionalism** with which the teacher's job[4] is done.

D. The **Ethics** with which the job is done.

Now, the merit of a teacher is not the only factor that must be taken into account in personnel decisions. There is also the **worth** of the teacher to the school or district. For example, teachers of Italian—however excellent—are of little worth to a school if demographic changes mean that there are no longer any students who wish to take that subject; but this is in no way a reflection on their merit as teachers. Worth is a system notion, merit is an personal one. Worth is extrinsic, situational; merit is intrinsic, professional. Since worth is so situation-specific, it cannot be exhaustively covered here.

It has been argued that the four categories above can be reduced in number in two ways. First, it has been suggested that Professionalism is dispensable; it is simply a means to an end, and that end is covered in the first two dimensions. Someone who taught marvellous material with enormous success could surely not be downgraded for ignoring all the 'rules of the game'. But the job of a teacher is not just teaching, and the Professionalism dimension picks up all the rest. Also, the 'rules of the game' in

---

[1] The *resources available*, not the *ability available*. The difference is crucial. If the ability is n't enough, that is no excuse; if the resources make it impossible to do the job, that is an excuse.

[2] Including the love of learning.

[3] Also including the love of learning; in this case, the extent of that love.

[4] All of it, not just the teachin' part.

this case are not just advice about the best way to do things, they are requirements as to how they should be done, so we have to have Professionalism in there to ensure justice in the way that classrooms are run. And finally, they are *also* likely to improve success. Since we are essentially never going to be able to pick up all the long-term learning that a teacher produces (the Quantity dimension), we have to settle for checking on short-term learning. But we can—and need to—buttress the bet that short-term learning will be a good indicator of long-term learning by looking at whether the approach is professional, *because a professional approach increases the chances of long-term effectiveness.* For example, sound test construction and marking—part of the professional repertoire—is more likely to pick up the places where further instruction is required, in time to provide that instruction.

Recapitulation But aren't style indicators based on exactly the same claim of improved long-term effectiveness? It is crucial for those using the duties-based approach to understand fully the difference between using criteria of professionalism and using style indicators, so we review the argument in slightly different form. Both are indeed indicators of success—at least, supposedly. The difference lies simply in their status as obligations; professionalism is an obligation, and it's a bonus that it also increases the chance of success. Style is quite different. Since it is clear that many teachers of the highest quality—perhaps the best teachers of all—do *not* use the style that research has been shown to be the most successful (for otherwise the correlations of that style with success would be 1.00), one cannot argue that the obligatory way to teach is to follow the winning style. But the best way to test involve setting valid tests, in each and every case, at least because justice requires it, and also because it leads to better diagnosis.

Ethical vs. Professional The second suggestion for compressing the list of basic criteria is very different. It begins with the suggestion that ethics is simply part of professionalism. The point is sound in principle, but we do normally distinguish between *codes of professional ethics*, which usually refer to ethical matters specific to a profession (in teaching, this would include the condemnation of taking bribes in return for giving high marks), and *recommended procedures* within a profession, such as the use of distracters in multiple-choice questions that are about equally attractive to a student who has not studied the materials.

However, there is a deeper analysis which links all four of the criteria into one, via the notion of duty. It is the duty of the professional teacher to ensure that the Quality requirement is met; that the Quantity learnt is as high as possible consistent with other duties; and that Ethical standards are met. Hence the system set out here is monolithically based on the notion of professional duty, although within that structure it incorporates the four dimensions already identified. At this point, we must therefore unpack the notion of duty into a more specific list of obligations; and then we must turn to the procedures for determining how well they are discharged.

17. The Professional Duties of a Teacher

Now we come to the heart of the matter. This is a long list and we turn later to the question whether it's too long. It's a list compiled on the basis of validity, not ease of measurement, and we come later to the question of how to measure these aspects of professionalism. And it's a list that avoids any reference to style or to any other indicators that cannot be shown to be *necessary consequences of the duties of the job*. In what follows, we describe each category of duties as a dimension, or sometimes as a criterion. The latter term is based on the logical use of the term, according to which criteria are what definitionally constitute an entity's essence, and the fact that these criteria (allegedly) define the teacher's job. (The distinction is between criteria and 'mere' indicators, which are only empirically linked with doing the job well.) But there is another use of 'criterion'; it is sometimes used to refer to the standards that must be met on each dimension. To avoid confusion, therefore, we use 'dimension' except when we need to distinguish criteria from indicators.

The sources on which this list is based include a large number of official documents which make some attempt at the same goal. A more important source has been the suggestions and reactions of several hundred experienced teachers and school administrators in Australia, the US, and Canada. As we proceeded through version after version, each new group was asked: "How would you, as a teacher, feel about being rated according to the way you perform the items on this list? What is missing that should be included: and what is included that should be omitted?"

At this point, the official documents we now see rarely score very well. They rarely include more than two-thirds of the items on the list; and they nearly always include a number of items that are not duties at all, or not duties in most school jurisdictions. The better ones we see now often derive from an earlier edition of this list, which has been widely circulated. However, that interpretations suggests that it's time for some more external criticism[1]. The order used here is not a presumed order of importance, but an approximation to the order of dependence; that is, the earlier items are usually required before the later ones can be handled.

1. KNOWLEDGE OF DUTIES

Includes knowledge of the law and regulations applying to schools in a district or a state as well as the expectations at a particular school (e.g., division of responsibility in team teaching situations; expectations of assistance with out-of-class activities such as syllabus design, materials selection, school projects, clubs and societies, special student reviews). Includes understanding of the curriculum requirements and the duties in the following list.

2. KNOWLEDGE OF SCHOOL AND COMMUNITY

[1] Please send in suggestions or criticisms of this or any other part of this article.

Includes an understanding of any special characteristics, background, or ideology of the school, its staff and students, and of its environment. This is part of the needs assessment for planning lessons and curricula (jobs available, languages spoken, family educational level) and of the resources inventory (parks, libraries, museums, tertiary institutions, factories) that should affect instructional planning. And it assists with determining what standards of teaching, and what expectations as to dress and conduct—in school and out—to adopt or protest (should homework be set; what grading standards are used; is a female swearing or a male earring daring?).

3. KNOWLEDGE OF SUBJECT MATTER

A. In the field(s) of special competence

Subject matter knowledge should be at least enough to ensure that appropriate materials can be selected or prepared, and explained, and that student understanding of them can be appropriately tested; and to ensure that most questions can be answered correctly. Where questions cannot be answered, it must be known where answers can be found quickly (this requirement of 'resource awareness' includes museums, art galleries, etc., as well as reference works). Suggested guidelines for minimum subject matter competence are—for high school teachers—two years of successful tertiary study of each subject taught, and one year of such study for primary teachers. A degree with a major in the subject should be expected where teaching of college preparatory courses is involved. Competency tests to ensure the continued presence of the equivalent level of knowledge are an obligation of the employer (and often also of the training institution), since: (i) even for recent graduates, a certificate from a credentialing institution cannot be counted on to provide that assurance, and (ii) for mid-career teachers, some knowledge and skills have evaporated or become outdated, (iii) other knowledge has been added to what is covered by that standard since they graduated, often representing a large part of the curriculum (earth studies) and sometimes representing most of it (computer studies).

B. In across-the-curriculum subjects like English, study skills, personal/vocational awareness, computer studies, etc.

While only a minimum level of competence is required, that includes a good tertiary level of literacy in writing, speaking and editing (which, for example, excludes nearly all spelling, punctuation and grammatical errors); a modest competence in the use of computers in the classroom; and similarly for the other areas. Some of these areas have been added to the obligations of teachers quite recently. With or without adequate inservice training, they become part of the obligations of the teacher. The task here is not to determine whether it is reasonable or unreasonable to include them, but only to determine what is now understood to be part of the obligations of a professional teacher.

4. INSTRUCTIONAL DESIGN

A. Course design *The teacher should be able to develop course plans from a knowledge of what is required by the local curriculum regulations and testing mileposts, together with information (which may have to be researched) about student ability/achievement levels, and available resources. Course plans typically include: a list of objectives or topics for lessons and terms (course outlines); activity, project, lab, library, homework, test and field trip descriptions located on a time-line; at an appropriate level for each class; and in a form adequate for use by a replacement teacher or supervisor. Versions of these may be provided to the class, if this is helpful rather than inhibitive of note-taking or inquiry skill development. (Note that no requirement is included for detailed lesson plans (behavioural objectives, activities in ten-minute segments, etc.) although these are sometimes a useful device, especially for beginning teachers.)*

B. Selection and creation of materials *(Applies to the extent that the teacher is allowed/required to select or add materials to those provided.) Teaching materials, selected or created to fit into the instructional plan above, should be current, correct, comprehensive, and—where possible—well-designed. They should, where possible, provide or include references, applications and enrichment resources as well as basic instructional assistance (unless this is covered by the text or other materials); where possible they should incorporate a variety of instructional and doctrinal approaches, for the benefit of students who respond better to an approach other than that provided by the teacher; alternative viewpoints should be presented fairly, so that students can consider the range of views; and there should be enough to supplement presentations by the instructor, visitors, trips, texts, etc.*

C. Competent use of material resources *Appropriate use of materials, library, computers, field trips, laboratory and specialist personnel (e.g., librarian, school psychologist). This use must demonstrate 'informed user' competencies. At the simplest level—chalkboard writing and overhead transparencies, and the writing and diagrams in paper handouts—must be readable, a test which many tertiary teachers would fail. (It is helpful to have explicit or implicit knowledge of the simple guidelines on number of words per line and lines per overhead that guarantee legibility in the average classroom.) Preferably, the teacher should be able to use those more complex audio-visual and computer technologies for which significant resources are available in the relevant teaching area. Systematic and objective evaluations of available materials by self or others should be used as the basis for selection. There is no absolute need to use media or specialists in order to do good teaching; but if they are available, and will significantly improve teaching the particular subject to these particular students, at a cost which is well below the benefits, the professional should be able to use them.*

D. Course and curriculum evaluation *The teacher (in, and out of, class) should be able to employ discussion, individual interviews, observations, questionnaires and testing—formal or informal—to gather and systematically record data for later analysis in order to get: (i) needs and ability assessments with respect to content, level, approach and pacing; (ii) information about the success of cur-*

riculum options and instruction. (These goals do not require individual test results, the need for which is covered below.)

E. Needs of special groups Knowledge of the needs of special groups that may be encountered is important, including the deaf and sight-impaired, blacks and Asians, non-native speakers, fast and slow learners.

F. Use of human resources The preceding efforts should be supplemented by involving specialist personnel (curriculum specialists, audio-visual and methods specialists) where appropriate.

5. GATHERING INFORMATION ABOUT STUDENT LEARNING

A. Testing skills As a basis for advice on student progress, to students and their advisers (and administrative authorities), the teacher must create or select, and administer, suitable tests (construed in the widest sense to include structured observation, project analysis, etc.). Tests should: match the content or skills covered in the teaching and required curriculum (including assigned out-of-class work) at the difficulty level appropriate for the class; be unambiguous; not be overcued; have one and only one correct answer when only one answer is allowed as correct; be answerable by a typical student who did the class work and homework but not by just any student; indicate the marks or relative importance of each question; relate to useful continuing and future competencies, in an interesting way where possible; allow the student to display creativity, understanding and the capacity to synthesize and evaluate—where possible and appropriate; be specific enough to provide evidence to guide counseling and modification of class materials where appropriate. To do this requires a minimum level of professional understanding of the advantages and disadvantages of testing in general and of various types of tests, including: multiple-choice, short and long answer, verbal and written, structured observation, interview, and project tests. The teacher should understand the difference between and be able to construct appropriate tests for summative, formative and diagnostic purposes; the difference between tests for ranking and for grading, and between norm-referenced and criterion-referenced tests; and should understand the use of matrix-sampling and item analysis. If multiple-choice tests are used, it must be understood how to construct them so as to measure higher-level cognitive skills (a feasible but rather difficult task). The construction of rating forms for feedback by students on teaching and teaching materials should also be well understood.

B. Grading knowledge (marking, scoring, rating, diagnosing) Must understand the difference between: holistic and analytic scoring and the advantages of each; the design and use of scoring keys ("rubrics"); the fallacy of the 'A for effort' approach; typical sizes of test-retest and interjudge differences; magnitude of test-anxiety effects; how to recognize serious learning disabilities, etc.

C. Grading process To the extent possible, this must be done so as to avoid bias, especially on essay-type questions by: using coded papers; marking question by question, rather than paper by paper;

changing the order in which papers are marked from question to question; re-marking early papers to pick up any drift of standards; using and improving a scoring key. The reasons for each of these procedures should be understood.

D. Grade allocation Grades should be awarded consistently (equal grades for equal quality/quantity of work); appropriately (no Bs or As for work that is merely satisfactory for students at that level, no Fs for work that is around the satisfactory level, etc.); and helpfully (on standards that relate to the needs of the students; on parts or aspects of work as well as on whole performance, when the test materials are being returned).

6. PROVIDING INFORMATION ABOUT STUDENT LEARNING

A. To each student

(i) On class performance The teacher should provide—in class or, when more appropriate, in writing or in private discussion—an indication of how the instructor thinks the student relates in quality/quantity of response (if the latter is required) to the standards expected and preferably also to the range of quality of peer responses, especially if there is any chance of misunderstanding by a student of his or her comparative or absolute level of performance.

(ii) On each test The teacher gives correct answers, explains the grading/marking standards, and the individual grades when necessary, comments on common errors, preferably distributes examples of fully worked good and bad answers with comments (not necessarily using real answers, or ones from the same class).

B. To the administration In the typical school context, the teacher must provide the administration with information about student performance on a regular and timely basis as required; must identify problem behaviour, and facility or support deficiencies; must call for assistance as necessary.

C. To parents, guardians and other appropriate authorities The teacher communicates to those with a right to know, and only to them, as to how the individual students or classes are progressing. Preferably, has the skills to enlist support from these people in the enterprise of motivating and assisting the students in learning.

7. CLASSROOM SKILLS

A. Communication skills The teacher must be able to communicate information, explanations, justifications, expectations, directions, and evaluations to students of the age and abilities that will be encountered in the place of employment. Success in communication requires efficiency and clarity in presentation and skill in the maintenance of attention. Competence in the engendering of motivation is desirable. Similar communication skills are required with respect to peers and parents, supervisors,

and sometimes community groups. Complete determination of this competence would depend on later outcome checks, but something can be picked up in the course of a classroom visit by a specially trained observer. (There remai.. the difficulties that the sample observed: is not random; is too small; is usually judgmentally assessed; and only refers to short-term success.)

B. Management skills

(i) Under emergency conditions Teachers have moral as well as legal responsibility for coping to the extent possible with what happens in an emergency. In particular, they should know what to do in case of any of the following that is possible in their area: (i) Fire; (ii) Flood; (iii) Tornado/Typhoon; (iv) Earthquake; (v) Volcanic eruption; (vi) Blizzard; (vii) Civil disorder (riots, tear gas, bombs, mob or strikers entry to classroom); (viii) Trauma, notably fractures, snakebite (or spider/scorpion bite), stab or gunshot wounds, electrocution, choking, gas-poisoning and seizures. (Five of the eight have occurred within the last decade in each of a number of metropolitan areas.) Field trips or overnight stays introduce other hazards such as the risk of drowning, which engenders the duty of mastering CPR techniques and the identification of poisonous plants, snakes, spiders, etc.

(ii) Under standard conditions Teachers must have the ability to control classroom behaviour so that learning is readily possible—and can be assisted—for all students at all times, while preserving principles of justice and avoiding excessively repressive conditions. Justice requires making clear what the rules and penalties for breaking them are, and enforcing them consistently. It should include the ability to cope with a range of useful class modes including whole-class and small-group discussions, questioning, question-answering and listening; it is desirable though possibly not essential to have the ability to achieve high time-on-task ratio; certainly it is important to have the skills to deal with student inquiries in such a way as to encourage the inquirer to further exploration. Lack of classroom control leads to disruption of the school and not just the classroom, either through direct (noise) impact or through the grapevine; and it leads to severe penalties for the students who are willing but unable to learn because it is occurring in or near their classroom. So it is rightly considered a minimum necessary condition for competence. But a quiet classroom is not necessarily a learning classroom, and evaluation systems that just reward silence are seriously flawed.

8. PERSONAL CHARACTERISTICS

A. Professional attitude The teacher should be able to accept criticism constructively unless the criticism is demonstrably invalid or redundant; should solicit critical evaluation of various aspects of job performance from time to time, including student evaluations where possible; should exhibit a positive attitude towards students and to teaching as a vocation; be helpful to parents, peers and administration with respect to legitimate requests; be helpful to apprentice or paraprofessional teachers; must not evidence prejudices related to race, religion, age, gender, etc.; must be punctual and conscientious

in performance of duties; must be compassionate as well as just in dealing with students; must, in general, be highly ethical in dealing with all job responsibilities and personnel; must try to avoid penalizing students in the course of industrial or personal disputes. Standards of language and deportment must be consistent with knowledge of possible impact on students.

Note (i): Being noticeably 'under the influence' of drugs such as alcohol while on duty is thus prima facie evidence of serious misconduct since it will affect capacity to perform the primary tasks, and probably affect respect for the individual and the staff in general, with consequent long-term costs in student learning. But being under the influence in the pub on Saturday night is part of the right to enjoy oneself in one's own way, as long as it doesn't interfere with the rights of others.

Note (ii): Leadership skills or achievements, often included as desirable for teachers, are completely inappropriate entries; they make a good basis for selecting future administrators, which is why they get mentioned, but they are entirely unnecessary for good teaching. The same applies to 'good at working in groups' unless it is a duty of the position that team-teaching be done. Some committee work is no doubt a common obligation, and should be rated on outcomes, not presumed components.

Note (iii): Counseling or 'pastoral care' skills would be appropriate for some jobs and not for others; the job description should be clear on this point.

Note (iv): There is a legal 'duty of care' meaning the duty to take care of students who are in your charge (especially when they are too young to do so without your help). Beyond this, it is arguable that a teacher should 'care about them'. But there is no duty to care for them as if they were your own, or even as if you liked them all. This is an area where well-meaning administrators often require more than is appropriate. It is crucial to professional service that 'distancing' be possible, or else the stress load becomes intolerable for many teachers we can ill afford to lose. The commitment declines with the age of the student, so that at the primary and secondary though not the tertiary level, it is important that teachers have a real concern for children's welfare, including their self-esteem. (This does not, however, entail a heavy-handed positive reinforcement strategy in primary school.)

Note (v): Enthusiasm for the subject matter cannot be justified as a requirement. A 'positive attitude' towards teaching, recommended above, is fully compatible with radical and sustained specific criticism of its condition and management. It is incompatible with unremittingand unconstructive denigration, which has a very serious effect on others, especially beginners.

B. Professional development Teachers should have good awareness of their own areas of strength and weakness and implement systematic procedures for self-evaluation and development where appropriate. This might include evaluation of their time and stress-management ability; engaging in systematic improvement of class materials and plans; experimenting with variations of method and/or materials to produce steady improvement; soliciting input from students and peers; engaging in systematic reading or other study of current developments in pedagogy and educational/text materials in

the teacher's area of specialization; being able to set out the results of the preceding efforts in a professional portfolio.

9. SERVICE TO THE PROFESSION

A. Knowledge about professional issues Without some knowledge about the profession, (its nature, role, history, current problems and issues), there can be little effective service to it. Without service to it, there is little of the profession about it.

B. Professional ethics Knowledge about (and performance in accordance with) the standards of the profession, e.g., in not representing oneself as presenting the school's viewpoint unless specifically empowered to do so. Acting so as to provide a good role-model for peers and trainees; perhaps assisting with activities such as the development and enforcement of professional ethical standards.

C. Helping beginners & peers Providing systematic assistance to beginners and student teachers should be regarded as part of the essential commitment to professionalism.

D. Work on projects for other professionals Examples include working on a newsletter or journal, organising a study group or making seminar arrangements, or working for a union. These would be appropriate though not mandatory.

18. Standards and Definitions

This is a long list and it would be unreasonable to expect a very high level of performance on every one of the dimensions. But it is contended that none can be entirely dismissed. A minimum level of achievement on every one is *required* and a substantial level of achievement on most is *expected.* Less would be *accepted* if the reasons for the exceptions were good, and only rarely invoked. Exactly what this means will vary in particular circumstances, and some case studies that illustrate limits must be included in any training workshops. It will be clear that merely adding up the score across all duties is an invalid integrative procedure; the minima must be achieved on each and a failure to meet them can't be traded off against over-minimum scores elsewhere. If the minima are met, totalling depends on weighting the dimensions. It is never easy to justify differential weights, but one might argue for halving the weights of items 1,2, and 9. But first, perhaps, we should look at the question whether there more fundamental problems with the list.

There is no doubt room for improvement of the duties list. That it will be in the nature of refinement rather than a radical alternative is a matter of the common linguistic use of the term 'teacher'. It is impossible for a user of the English language to suppose that classroom teachers have few of the above duties; certainly the main thrust of the middle five items (3-7) and probably item 8 is part of the

everyday concept. The details of these items, and the other items, emerge as part of what experienced teachers and school administrators feel are duties of the profession.

With the preceding remarks, we have concluded an answer to the question which people often naively think must be answered *before* any study of the evaluation of teaching can be done: the question of how to define good teaching. The real situation is that one works towards it throughout a study of the field; it is a major *goal* of serious research, not a minor preliminary to it. Good teaching is whatever scores well on the duties list, with the provisos just mentioned (and some attention to the next few sections as well). Of course, this is messy compared to a one-liner, but one-liners are jokes. Or, at best, convenient but inaccurate mnemonics; they are never full encapsulations of the meaning of major concepts, in physics no more than in politics.

The test of whether some factor is part of the *definition* of 'good teaching' must be distinguished from the question whether teachers at a particular moment in history think it is part of *preferred practice* in teaching. Preferred practices are bets about what works in achieving whatever the definition of good teaching requires. They are not part of the meaning of the term. Yet we often find studies where the distinction just made is confused, and it is suggested that practices which are excellent in some areas and irrelevant to other areas are part of the very meaning of good teaching, e.g. highly organized presentations. Socrates would have failed on this criterion, which is a counter-example to the view that it's part of the definition. Organization, like eloquent speech and evident enthusiasm for the subject-matter are style variables. A skilled *helper* can recommend them, with care and some risk; there are reasons to consider doing so *in some cases*. A competent *judge* cannot use those standards. The formative context is *critically* different from the summative.

19. Rules of Evidence

Instead of going into step-by-step details on documentation, which would run us beyond the space available, some general comments are made which will indicate how performance on each dimension can be evaluated. It must be noted that these suggestions are validity-oriented, not politics-oriented. The comment made earlier about political compromises applies here.

A. In general, more than one source of evidence—and preferably more than two—should be used to document performance on each criterion. These matters are all prone to some inaccuracy, many of them prone to understandable bias, and their importance demands the use of confirmatory evidence. Where possible, the sources should be independent, that is, not subject to effects from the other sources.

B. It is essential that in all cases except where overwhelming considerations of confidentiality apply, the teacher—like the reviewers—has the opportunity to see and respond to all evidence. This

immediately provides a second source of data on most dimensions, though not a fully independent one.

C. Where major or crucial disagreements occur between the estimates from different sources, further investigation must be undertaken whenever there is any chance that it will resolve the issue.

D. In no case is automatic averaging of estimates justified. The person in charge of personnel decisions is responsible for making decisions about which sources of data or judgement to take into account at all, and what weights to assign to testimony or dimensions[1]. (There will of course be some individual judgements or dimensions in some individual cases where averaging is the best policy, but the evaluator should make and be responsible for that decision.) In turn, the principal component in the evaluation of the evaluator is the care and skill with which the evaluation of teachers is performed.

E. Significant, and—especially—systematic, inaccuracy on the part of any estimator[2], as revealed by comparing their ratings with objective data or with the ratings of others, is sometimes a sign of professional incompetence in itself and should be rated as such. If the teacher, for example, constantly overrates his or her own performance, doing so is a sign of poor self-evaluation skills, and these are part of the set of required skills for any professional (Dimension 8B in the above list, for example). To an even greater extent, since this skill represents a larger proportion of the relevant professional repertoire, the personnel evaluator is vulnerable to criticism for biassed or inaccurate ratings, which will automatically be entered into his or her file.

F. Systematic inaccuracy should also be used as a basis for extrapolating with a correction factor, so as to ensure that no-one suffers or benefits from persistent bias.

G. In view of the above, it is very important that teachers collect as much documentation relating to their discharge of duties as is possible, so as to reduce reliance on judgement or memory. This 'teacher portfolio', for which they are entirely responsible[3], should be part of their official personnel file, used whenever reviews occur, and work on it should be supported by appropriate inservice workshops. Providing such documentation is part of the professional skill repertoire, partly because it is relevant to good self-evaluation, and partly because it is a contribution to the effective governance of schools.

H. An appeal process, for all parties, against any decision, is essential. It is not essential that it be a massive and exhaustive one, and almost all appeals should be managed by an arbitrator whose sec-

---

[1] Weights only come in for rewards, since no weighting can, in general, offset a failure to achieve minimum standards. Deviations from unitary weighting are always very hard to justify. but should be determined in advance and announced to applicants.

[2] The estimators are the sub-evaluators, those whose ratings are being combined by the evaluator in order to achieve the overall evaluation.

[3] Apart from annotations from reviewers, which have been seen by the teacher, and to which the teacher's responses may be appended.

ondary commitment is to develop a set of guiding principles and case law that will reduce the necessity for large numbers of appeals. Appeals should not be heard unless they raise a serious question, and appealing without appropriate grounds is itself, in some cases, unprofessional conduct.

I. Evaluation systems like the one proposed here are complex and should never be attempted as one-shot implementations. Stage I should involve a one-year test, for volunteers only, under a no-harm guarantee, and should lead to significant refinements. Stage II should be a two- or three-year trial with serious external evaluation.

J. Following the example of several school districts, the job of the school principal, if s/he is to be the main evaluator, must be redefined so that staff evaluation and development are held to occupy the top 25% of the workstack; that is, nothing short of safety emergencies should pre-empt this commitment. They should probably be treated as worth 50% of the points in the evaluation of the administrator (because their worth to the district is so high).

20. Sources of Evidence

"Evidence" refers, at least, to the following: expert testimony *in the area of demonstrable expertise*; 'found data' (existing records); incidental or specially arranged observations; and to the results of tests and experiments. The way to set up a systematic approach here is by means of a matrix which shows how each of the duties will be covered by data from two or three of the sources listed below. Exactly how that matrix will look depends on the local situation (are there assistant principals, what does the contract allow, how many teachers have to be reviewed, etc.)

A. Judgements By: the teacher engaged in the evaluation, other teachers, department heads, counsellors, students, parents, principals, district personnel, inservice providers, etc.[1]

B. Found data School records: e. g., applications for this teacher's classes, attendance, grade distribution, recommended texts, student work, tests, class handouts, assignments. Teacher records: lesson plans, log of notes on students, on classes, on success of materials, etc. Personnel records: original job description, letters of support and complaint, applications for transfer, enrolment and grade records, etc. Library records: assignments, checkouts, etc.

C. Observations In the classroom (constrained to duties only), in the school grounds, in the common room, in committee meetings, in dealing with parents/peers/students/counsellors; done by students, peers, administrators.

---

[1] Of course, there is no suggestion that, for example, fellow-teachers be asked for an across-the-board evaluation of of the teacher. They are not in any position to give any such evaluation. On the other hand, teachers with expertise in the same subjects would be good judges of materials and grading standards.

D. Test data Done on the students of this teacher, or in which the teacher participated. On: comparative performance, absolute performance. success of particular approaches , materials, teams, etc. Or done by others, with this teacher as a subject, e.g. state competency tests.

E. Teacher portfolio Self-evaluation and personal development plan; results of experiments and reading program; courses taken; procedures used in grading; basis for selection of materials.

F. Footprint data The results of exit interviews with graduating seniors are often valuable additions to the data; college applications/acceptances/scholarships; changes in the curriculum due to this individual's committee work.

21. Advantages of the Duties-Based Approach

In the light of this fairly detailed picture of the DBTE model, it may be appropriate to summarise its claimed advantages. Perhaps the most important is that it avoids the use of illicit material and inappropriate judges, thereby reducing the chances of injustice and large legal and damages costs. In particular, it places very strict limitations on what can be picked up from classroom observation. Second, it brings in a good many factors that are normally overlooked although they are of very great importance, ranging from the quality of content to performance on committees. This feature of DBTE means that it gives a better picture of the teacher's total contribution, including the out-of-classroom contribution, and it makes the teacher feel that these dimensions of performance are appreciated. Third, DBTE brings in sources of evidence that are often ignored, most notably the teacher portfolio and the student ratings, but also the footprint data, which increases the extent to which long-term student benefits are given weight. And there is an overall improvement in the solidity of the evidence, and—though this is a judgement based only on anecdotal evidence—in the sense of participation (reduction of alienation) on the part of the teacher and the students, whose input is so important. Fourth, it encourages rigid segregation of the formative from the summative personnel, so as to improve the rights of teachers to be judged for what they do, not judged on whose instructions they follow. This segregation also improves the chances that the support system will be utilized.

22. The Teacher Development System

The general thrust of inservices should be driven by the results of the staff evaluations. To take one example, it seems clear that one of the most serious deficiencies in teacher preparation and practice, from primary through tertiary, is in test-construction skills. This is highly teachable, and might be built into a series of inservices required of those who do not pass a pre-test; it's generally thought that it should be joined by computer-user skills for the substantial number of teachers who have not yet acquired them.

One might suppose that this, at last, is the area where one can use all the research into the relative effectiveness of different approaches. Unfortunately, the matter is not that simple. The research does have a place here, but one that has to be very carefully circumscribed. (Even in preservice training, it can only be used within strict limits.)

It's best to distinguish two kinds of situation from which the teacher comes to inservice. In the 'standard updating' inservices, demonstrations of and workshops on research-based approaches are desirable as long as they are not *required* of all staff. One must bear in mind that what improves one teacher may cause damage to another. We know from the statistics in the better studies on Direct Instruction, for example, that teachers tend to improve their performance, measured rather simplistically, on material which suits that approach, at least for students in certain age ranges. What we don't know is how many teachers of material quite different from that on which the DI approach has been validated would have their performance damaged by adopting this approach (e. g., teachers of moral education or literature or current affairs). In fact, we don't even know whether the ones that show the short-term gain deteriorate below their initial performance in the long run, because there are no comprehensive long-run studies around yet. When playing with fire, only use volunteers—and you can still get sued, as the Army found with Agent Orange. If attendance at the inservices is compulsory, you must offer more than one parallel activity. And if one of the options is a workshop on method, make sure the alternative to it has nothing to do with method, or you will find yourself with a potential responsibility for damage that is only diluted, not removed.

There's a second type of teacher background to inservice which calls for more desperate measures. This is the 'last chance' situation, when a teacher has been identified as unsatisfactory and must show improvement (not the same as 'must attend specified remedial exercises and do the assigments required by them') or else lose job or status. In a last chance situation, the helper should indeed recommend to the teacher—remember, it's still the teacher's option as a professional to accept or reject advice on remediation[1]—the best research-based approach available. For two things are known that were not known in the 'regular updating' situation. It is known that this teacher's 'natural style' doesn't work; and it is known that the risk of having their teaching modified for the worse is not the most serious risk they now face. And Direct Instruction (aka Precision Teaching, etc.) is probably the best bet around, for areas related to those for which it has been validated and for teachers whose alternatives are known not to work satisfactorily.

23. Mentors

Various people can serve in the role of the 'helper' mentioned in the previous section. One important possibility that also addresses the need to extend the upper level of the salary scale for good teachers

[1] Otherwise, if the advice does not work, the person who gives it is at least a co-author of the failure.

is the 'mentor' position. In fact, the mentor position, defined in some such way as the following is almost an essential creation if a district is to show that it really does value teaching. (i) Instead of the usual revealing arrangement in which the best salaries in the district go only to administrators, the mentor is selected simply on teaching merit and is paid on a salary that runs up through that of principals of smaller schools. (ii) There are only a few mentors in a district (the suggestion is about 1 per 50-100 teachers), so that the job is prestigious and not too expensive for the district.

(iii) The appointment, although renewable, has to be re-earned in open competition every three years; it is an *award*—a salary increase against increased expectations in the future—not a *reward*, something given in recognition of past performance. It would make a travesty of the position to assume that superior teaching skills—any more than competent teaching skills—do not deteriorate, since one would soon have mentors who had burnt-out or stopped keeping their teaching up-to-date once they got the award. (iv) The only requirements of the position are to allow other teachers (including student teachers) visiting privileges—one at a time—to the mentor's classroom, and to be willing to talk about teaching *informally* and *occasionally* with other teachers at the same school.

(iv) An *option* of the job is to do some formal helping of other teachers who request it, up to 40% of time, if that can be accommodated by the time-table and the district funding. It can't be made a *requirement* of the job, or your new way to reward good teachers turns out to require that they take on teacher-training work instead of administration. This is a partial replay of providing an incentive for good teachers to get out of teaching. In which case, they should be selected at least partly for that skill, which means not simply for teaching skill. Similarly, professed willingness to undertake the larger role cannot be taken into account in selecting mentors, or in reappointing them. The mentor must be seen as primarily a super-teacher whose benefits are primarily to the students. They can only guarantee that they will be role-models, not that they will be good teachers of teachers. Apart from anything else, we know from other areas that the best swimming coaches sometimes can't swim at all, let alone well; and that the best football players are often hopeless as coaches. We should not build the mentor system on the opposite assumption. (v) Mentor help may be made available to staff from nearby schools, if the need at the mentor's own school is met first, and the mentor, principals and district approve. (vi) It is unlikely—and it is not essential—that the school system will cover the time of the teachers who take advantage of the help as extra released time. It can be covered under regular inservice time or as an 'after hours' activity. (vii) The 'helping option' is compensated for with released time for the mentor, and (viii) performance as a helper is largely evaluated by those who use it, though it may be monitored for accountability purposes.

(ix) If the mentor does exercise the formal helping option, then a confidentiality requirement applies. It would be ideal if no information about the occurrence, extent or results of helping interactions that identifies any participant were to be passed on by the mentor to anyone else. But accountability must apply to the mentor as well as to teachers, and it makes the complete ban impossible. Hence, the

names of those getting assistance—which might of course be assistance in going from good to excellent—must be provided to the principal, along with an indication of the time spent with each, so that the principal can give and collect from these individuals their rating of the mentor. What cannot be passed on is any suggestion as to the mentor's view of their merit, or their diligence. A hearing for instant loss of mentor ranking, and loss of job, follows upon a breach of that condition. (x) Mentors should probably have someone in the district or at least the state office to whom they can turn for assistance with their helping role (for example, to get copies of materials, providers of courses at tertiary institutions in the area, etc. (xi) Mentor positions should normally be associated with schools, not be free-floating, though a mix of the two is possible. Otherwise the mentors are likely exercise seniority rights to congregate in the most-favored school in the district, abandoning the schools with the most need of them, as in the Australian Capital Territory implementation. (xii) Teachers at the school, and any willing to transfer to it, must both be eligible to apply. It is desirable if no preference should be given to those from the school, since their better knowledge of local conditions is offset by their lack of independence and lack of familiarity with alternative approaches. This avoids 'closed shop' expectations growing up into a norm which makes it very difficult to bring in new blood. (xi) The normal arrangement would be that a mentor is appointed when a teaching vacancy occurs and the principal and/or teachers, as well as the district, feel that it would be desirable to have a mentor at that site.

24. Worth

Teachers have other values for a school besides professional competence. Being well known and liked in a district because they grew up there is likely to be valuable for school-community relations, although it has nothing to do with professional qualifications. Versatility in academic coverage is another example of a kind of value to the school that transcends the performance of immediate present duties; it lies in the area of potentiality. But there are other qualities that a particular school, in certain circumstances, may rightly value, including gender and religious conviction. In other situations it would be quite improper to weight these. This is a matter of looking at the mission statement and the school's demonstrable needs in some detail, not a matter for simplistic slogans that suggest gender and religion are always inappropriate considerations. Distinguishing between proper and improper uses of 'worth' factors requires some extensive discussion of particular examples.

25. Conclusion

We have reviewed a number of feasible approaches to overall teacher evaluation. It seems clear that only one of these is valid. But it is relatively untried. What should the sensible school administrator do at this point? What should teachers support and encourage?

It's a good general principle in educational administration to let others play guinea-pig. Moving only to the 'tried and true' avoids wasting effort on debugging new approaches, and quite often the fashion proves a flash in the pan so that the need to change evaporates. The situation is different here. Although we cannot point to long track records with the proposed system as a totality, there is nothing unfamiliar about either its data sources or the duties to which it appeals for validity. In fact, we have good evidence that the administrative infrastructure for this kind of approach is perfectly workable, from the sites where Ed Bridges' approach (outlined in ***Managing the Incompetent Teacher***, Stanford, 1985) has been implemented. And the alternative is to continue using a system which is demonstrably unjust and almost certain to incur expensive penalties.

From the teacher's point of view, it is surely preferable to work with a system of appraisal that does not change with every new batch of research results, and that does fully recognizes the great range of a teacher's duties. The appropriate response to this situation, for everyone, must surely be to move as quickly as possible to eliminate every use of style criteria or impressionistic judgement from existing practice, replacing them with duty-related data with its the many missing types and sources of data.

When you have discovered an uncontrallable fire in the classroom, it is not appropriate to hesitate about taking the students to a new building on the grounds that it lacks a long occupancy record.