1869

More Accurate Evalutaion of Student Performance.pdf

Peter Elbow

More Accurate Evaluation of Student Performance

BY PETER H. ELBOW

It's not a question of whether we like evaluation. When a teacher sees student work he almost invariably has an evaluative reaction. Even if he doesn't, the student almost invariably infers one. Even tone of voice and facial expression play a role here. Besides, we couldn't learn without feedback. Therefore, the only real question is what sort of evaluation to have. We decide best if we figure out what evaluation ought to do.

There are two purposes. The first is to provide the audience with an accurate evaluation of the student's performance. If the student or some other justifiable reader gets an inaccurate impression, the evaluation has failed.

The second function of evaluation is to help the student to the condition where he can evaluate his own performance accurately: teacher grades should wither away in importance if not in fact. We haven't fully taught someone to do something or know something unless he can determine on his own whether he has done it or knows it. A student who remains dependent on the teacher's grades for evaluation is defectively taught in a simple, functional sense: he cannot, strictly speaking, do

PETER H. ELBOW *is a lecturer in literature at the Massachusetts Institute of Technology.*

what he was supposedly taught to do because he cannot do it alone; he cannot do it unless someone simulates for him the old conditions of learning.

We see here that the program for grading reflects what seems to be the program in many cognitive activities: the organism must learn to make internal and autonomous an activity that originates as interaction with something outside itself.

It might seem at first as though the two functions of grading—to evaluate accurately and to wither away—are at cross purposes. But actually they work together. The best hope for teaching trustworthy self-evaluation is to give a more accurate and explicit message of evaluation than traditional grades contain. Grades can only wither away in importance when they cease to be ambiguous and magical. The present system too often allows the student to feel them as judgments based on hidden criteria, judgments which he cannot understand and has little power over. If he is rewarded he feels he did the right things, but if the reward fails he never knows which step in the rain dance he missed.

Both functions of grading can only be served if we confront the central question: What constitutes good student performance? Other mooted issues—are grades necessary? do they harm? should the student see them? should they be quantitative? how much should they count?—are really ways of avoiding this question.

We can make headway on this question if we begin a catalogue of components of good student performance which teachers actually imply in their grades—the various messages or definitions of student performance that various teachers consciously or unconsciously imply:

—Command of course information.

—Memory.

—Understanding of the central concepts of the course. (Note that a student may do well here without producing a lot of information or seeming to have a good memory.)

—Logical, conceptual intelligence.

—Application of the central concepts of the course to new instances, seeing the concepts from new and creative points of view, seeing new implications. (Note again that this capacity does not necessarily imply the earlier ones: it can accompany a bad memory and serious misconceptions.)

—Creativity, imagination, intuitive insight.

—Effectiveness of verbal strategy: How good is the student's communicating or rhetorical skill? What is the proportion of message to noise? (Noise comes not only from unclearness and awkwardness, but from anything which obstructs: inappropriate syntax, spelling, mannerisms, and so on.)

—Effectiveness of thought strategy: How well does he come to grips with the question or formulate the question behind the question? Does he see to the heart of the real issue and deal with it persuasively, or does he spend too much time saying things which may be true but are not the strongest way to satisfy the question? (Needless to say, the distinction between verbal and thought strategy is rough and problematical.)

—Curiosity.

—Permanence of learning.

—Integration of course matter with what he already knows.

—Growth or improvement.

—Utilization of potentiality.

—Potentiality for further development.

—Judgment.

—Diligence, effort.

—Moral trustworthiness.

—Likableness.

—Enjoyment of learning.

There is an easier way to go about categorizing student performance: performance on papers, in laboratories, on examinations; attendance; preparation for class; participation in class; work on time in acceptable form. But this begs the question of what good student performance is.

If a teacher scorns some of my earlier entries, let him investigate more fully the grading behavior of some of his colleagues, or

himself. He will discover my list more parsimonious than wild. It would not be difficult, for example, to show that even in college some teachers include dress, appearance, and carriage in their grading, and not merely as accidental corollaries of other factors.

In addition to the terrific diversity of components that a grade is likely to imply in the hands of different teachers, the meaning of grades is further complicated by the fact that the same teacher is apt to treat a component differently at different ends of the *A*-to-*F* continuum—for example, to allow diligence, memory, or improvement to operate at the lower end of the scale and not at the upper end.

A slightly different phenomenological logic is not uncommon: *A* and *F* are for performances causing acute pleasure or pain, a powerful jolt for the teacher one way or the other of surprise, insight, excitement, or anger, disappointment, disgust; *B* and *D* are for performances yielding definite satisfaction or disappointment; *C* is for the affectless middle. Anyone who pretends to be shocked that grades should measure the affective response of the teacher ought to direct his energies instead to what is genuinely shocking: that it is so seldom admitted. What we need are methods either for preventing the activity or for letting it be clearly admitted and explained, thus sharpening the effectiveness of a tool which undoubtedly can be far more perceptive and acute than purely cognitive discrimination.

This partial analysis of messages implied in grades will serve to suggest more theoretical questions: Is the grade a measure of a particular performance or is it a statement about the characteristics of a person; that is, does it mean, "He remembered X quantity of material today," or, "He has a memory of X quality?" The former can be called the only warranted message. But it can also be called evasive.

Implied here is the question whether grades are a measure of past performance or a prediction of future performance. Since inferences about future performance are bound to be made, the operational question is who should properly make them. It can be well argued that certainly the teacher should not; it is beyond his province and hence unfair. But it can also be

argued just as well that since someone is going to do it, he should, since he knows (or should know) more about the validity of the test on which the past performance is based.

Furthermore, there is the question of what the individual student performance is to be measured against. The class? The school? The nation? The student's potential best? Or is there some standard implicit in the subject matter itself?

Needless to say, these questions do not admit of easy answer. But if asked, they admit a few tentative agreements and many shared and articulated disagreements. Unasked, they admit only hidden ambiguity, inaccuracy, and misunderstanding. If this central and difficult question of what is being evaluated can be squarely faced and dealt with, even if not neatly solved, most other issues about grading can be satisfactorily worked out.

The crucial conclusion is obvious: there is no need to have only one factor in a grade. There is no reason why a university, a division, or a department cannot come to agree on a grid of five to ten factors among which any teacher may choose. To illustrate the proposal, here is a grading grid with a conceivable set of factors. I am not proposing them, nor suggesting that the previous catalogue suffices as a list to choose from.

Name: _____ Pass ☐ Fail ☐

	(weak)		(strong)	
1.	☐	☐	☐	Command of course information
2.	☐	☐	☐	Understanding of central ideas
3.	☐	☐	☐	Imaginative and creative use of subject matter
4.	☐	☐	☐	Verbal strategy
5.	☐	☐	☐	Thought strategy
6.	☐	☐	☐	Class contribution (preparation, attendance, participation)
7.	☐	☐	☐	Growth over semester
8.	☐	☐	☐	Diligence, effort

The value of such a system would be in its flexibility. (One of the categories could even be the traditional *A*-through-*F* continuum if some teachers felt they could not accept a different system.) Any teacher could use as few or as many factors as he thought proper. Perhaps one man thinks the first factor is the only proper one. Fine. But let him admit it, and also permit some of his colleagues to slice the pie differently. Indeed there is no reason why a teacher shouldn't use different factors for different courses, or for different students in one course. A student might happen to display a particular quality (or absence of it), such as diligence, and thus be evaluated on it (if the teacher thought it important). Yet it would be wrong to evaluate all his students on diligence unless he actually built in procedures to test it. He simply wouldn't know whether most of his students are diligent or not.

Probably most teachers will have two or three factors they feel are crucial, and will evaluate every student on the basis of them: papers and examinations will be designed to test them. Some of these teachers will feel it is wrong ever to check any other categories. Others will feel it is right to use additional categories when appropriate to a particular student. Some teachers, however, will not call any factors indispensable, but will merely use whichever seem most appropriate to each student's relationship to the particular subject matter. In short, the system's flexibility would allow evaluation to be more closely functional with the measuring instrument (the teacher and his course material) and the things measured (the individual student performance). To the degree that evaluation departs from those two things it is false and untrustworthy.

Notice that there would be no need to assume that all factors utilized had equal weight. The teacher will have his idea of what the relative importance of each should be, but why should he force this judgment upon readers of his grades? If he decides the student should fail, his reasons are likely to be clear, certainly more clear than with conventional grades. And if the student passes, who cares (in this context) whether the teacher thinks creativity is more important than memory

or the other way around? The whole point of this system is to let the teacher provide substantive information and force interpreters to assign their own values.

But how could a department, much less a university, ever come to agree on a slate of five to ten factors? Again a solution suggests itself if we ask the central question, namely, should the slate be the factors teachers do use or the ones they should use?

The two principles can productively interact. First an experimental semester. A committee would poll its colleagues and its ingenuity to make an exhaustive list of factors that actually are implied by teachers. This list, phrased concisely, could fit on one sheet of paper. For the experimental semester, teachers would use this long list for grading, with complete freedom to use as few or as many factors as seemed right. But the object would be for everyone to try to feel out all the factors and see which ones seemed valid and meaningful—to try out reality in terms of various schemes for conceptualizing it. (The process would probably suggest new categories or groupings which could be added.) Conventional grades might be given that term for official use.

On the basis of this experiment, a faculty could decide on a list of less than ten. A particularly empirical-minded community might be content simply to subject the results to factor-analysis on their computer to see which were most used and where the cut-off fell most naturally. But probably it would be better to start with the results simply as evidence, and on the basis of this and of everyone's experience in trying out categories, consciously debate and decide which factors ought to be used. Ingenious rephrasings and judicious amalgamations of categories would be appropriate in this process. The goal is to achieve the most economical set of terms for the richest disagreement. The debate would be heated, but it is the sort of debate that enlightens. It would force greater communication between disciplines and improve the spirit of teaching.

Grading during the experimental semester would be a bit more trouble, though someone would be sure to call down a

shower of money from a foundation for the pains. But the new system that emerged would be less trouble than the present one. It is the present system's indeterminacy and ambiguity that cause agony and long periods of indecision in figuring out a grade. Surely it would be less trouble, and even quicker with practice, to make clearly differentiated and defined judgments than endure the present headache of always having to subtract apples from pears to arrive at one quantitative result. Also it would suffice in the new system to have only three or four points on the continuum for each factor, instead of the conventional six of *A* through *F* (or thirteen, counting pluses and minuses).

Nor is the plan unworkable. We can simply ask the defenders of traditional grades why there is any necessity for summing up student performance on one scale so that the student body can be ranked quantitatively along one dimension. Even Selective Service no longer cares. Is there any reason why universities must satisfy the conditioned desires of various outside groups—employers, government agencies, and other universities—to know where a student ranks along one dimension? Particularly if that one dimension be judged specious? Under the new model, on the other hand, the university would be able to satisfy the more defensible desires of such organizations—the desire to know the strengths and weaknesses of a student's academic performance. The interpreter would have to make up his own mind about which qualities he is looking for. And if he is looking for some factor which the teacher didn't use, perhaps creativity or diligence, that would be a far better state of affairs than the present one in which conventional grades are used and the interpreter is liable to infer erroneously that creativity or diligence is measured. Perhaps the system would cause a bit more trouble to admissions committees of graduate departments, but every teacher knows, because of the growing need for letters of recommendation, that there is little real trust in the meaning of present grades and class rankings. (Letters of recommendation are often vague and difficult to assess. The discriminations that would turn up

on the proposed system are just the sort needed in such letters.)

On the other hand, we can ask the attackers of traditional grades whether it would really be so bad to make quantitative discriminations between students with respect to one factor or another, so long as the process does not involve the mistake of summing up a student's whole performance on one scale and pretending you can measure all student performances on it. And if a student's whole performance is not summed up in one quantity, he is much less liable, indeed less able, to make the mistake of grounding his sense of worth in the teacher's evaluation. The evils ascribed to quantification would be minimized. "Hey! What did Jones give you in Nineteenth Century?" The question becomes considerably more complex. It could no longer be shouted on the run. A grade would be less often confused with a gift. The question would have to be seen more accurately or else disappear.

Even the registrar's office could handle this system. A student's four-year career could still fit on a single sheet of paper—thirty-two or forty little checked grids and a key. If the office has a non-refundable computing machine, all kinds of complex computations could be made on the basis of the various factors checked. Most of the results would be untrustworthy, but far less so than the computations on the basis of present grades.

Some will say the system might work in the case of a small class where the teacher knows the student well, but not otherwise. But consider the opposite conditions: a university which asks a graduate student to determine an unknown student's grade on the basis of only one paper and one examination. It is in just such cases that the traditional system is most unsatisfactory and the proposed one most necessary. The less data there are for making an evaluation and the more crude the instrument, the more necessary it is that the factors being evaluated be precisely defined.

A young, inexperienced graduate student is likely to be best at teaching and worst at evaluating. Good evaluation most requires experience and perspective and these are what the graduate student is apt to lack. On the other hand, the senior

professor is likely to be best at evaluation and—at times, unfortunately—worst at teaching. Thus the profound badness of some bad courses: everyone is awarded his worst role.

I hear a mathematics teacher saying, "Why all these categories? I teach mathematics! The grades I give are the sum of clear and unambiguous tests on mathematics!" Such a man could easily use the one category that fits best, perhaps "understanding of central concepts" or "effectiveness on examinations." But one could fairly say to him that if he cannot distinguish between the different cognitive or heuristic ingredients of his examinations, he proves he is no teacher of mathematics, however skillful he may be at computing correct answers.

I hear a tough man saying, "I refuse to let my university prostitute itself by officially sanctioning 'effort' as a meaningful educational category for college students!" But the important point here is that the present system does just what he objects to. It gives official sanction to whatever category blows across the fancy of every teacher, without the slightest need to make it conscious or articulate, much less justify it. Thus the proposed model should really offend not so much the tough man as the tender man who celebrates the present system because it allows total freedom and total diversity of categories. For the proposal does indeed limit freedom and diversity, but only to bring them within the limits of communicability. Celebrating the flexibility of the present grading system is like celebrating the flexibility of a radically impoverished language, such as a very limited slang: it feels perfect because it expresses and means every nuance you intend, but only to you, not to your audience.

It will be objected finally, and most damningly, that what I propose as an experiment is really a regression. The troops in the vanguard are conquering under the banner of Less Grading—note all the pass-fail experiments in progress—and here I come proposing in effect More Grading. But this brings us back to the functions of grading. I certainly want to be up front with the swingers, but I would try to clarify the inscription on the banner: Less Grading is only valid if it really signifies

the gradual transfer of effective evaluation from the teacher to the student himself. Pass-fail systems can potentially serve as a giant step in that direction; no grading, perhaps even more so. But self-evaluation is not easy, and unless it can be assured that teachers will talk regularly with all their students and comment copiously on their papers, at least in a student's first year or two, it seems important to provide models and processes to help the student learn to evaluate his work accurately.

This proposal for grading might lead some colleges or universities to other experiments.

First, a faculty which takes majoring particularly seriously and which is confident of its stature in the academic world might adopt the following plan: each department or division makes its own grading grid; the student receives the results of such grading for every course he takes, but his permanent record retains these results only for courses in his major or division; all other courses are either blank or pass-fail.

Second, students might be asked to evaluate their own performances for the last two or three years of college. Teacher evaluation in terms of clearly defined factors would prepare them to do this responsibly and accurately. Under such a system perhaps student evaluation would be reviewed by the teacher who would tell the student where he disagreed. Discussion might correct a misperception on one side or the other. The student could make changes in his self-evaluation if he wished, but he would never have to. His judgment would be final and official. He will only achieve really valid judgment in such difficult matters if he knows he has full responsibility and it is not just a game. (Of course, many colleges already have successful systems of official self-evaluation.)

Third, perhaps students should play an important role in determining what categories should be used in grading.

Fourth, sustained attention to the question of what is good performance will make many teachers wonder about the validity of "pass" and "fail" as categories: whether or not they are substantively meaningful once there is more than one dimension

to the grade. Some colleges with faith in the worth of their instruction and their students will dispense with these categories.

Fifth, the system might serve for some colleges with small classes as a transition to the use of only written—totally non-quantitative—grading.

But I would leave the emphasis not on the plans the model suggests to me but rather on the generative process itself—a faculty confronting the three problems in grading: (1) what constitutes good student performance? (2) how do you communicate evaluation? (3) how do you produce in the student the ability to evaluate his own work? If a faculty will sit down together to this task in good faith and with the sense that *some* solution is desperately needed, then whatever plan it produces should be right for it. In addition, of course, it will profoundly renew the spirit of the university as an institution for teaching.