2012

"Good Enough Evaluation: When is it Feasible and When Is Evaluation Not Worth Having?"

Peter Elbow

17

GOOD ENOUGH EVALUATION
WHEN IS IT FEASIBLE AND WHEN IS EVALUATION NOT WORTH HAVING?

Peter Elbow
University of Massachusetts at Amherst

Edward White's approach to assessment has always been practical and realistic. Over and over he has urged members of our profession to get involved in assessment, even as amateurs, to be willing to "get our hands dirty." Otherwise, he warned, assessment will be taken over by bureaucrats who know even less about assessment—or who would give the job to professionals in assessment who "know everything there is to know"—except what really matters. I like to think of him as good at playing both the doubting game and believing game with assessment. That is what I am trying to do here.

What I admire about Ed's work—and what I am trying to emulate here—is his constant attempt to do a kind of pragmatic realistic calculation: comparing the *need* for some evaluation (including how much information and precision is needed), and the *harm* or *risk* of untrustworthy results. If the need is great enough and the harm is small enough, then it makes sense to go ahead with it. This is what I mean by "good enough" evaluation.

He demonstrates this ability impressively in his ongoing thinking about holistically timed essays for placement testing. Famously, he came out in favor of them, saying they had shortcomings but were "good enough" given the need (White, 1995). But after a number of years of thinking, observing,

and conversing, he changed his mind and wrote this in 2007: "I have at long last lost confidence in placement testing as an appropriate method . ..." White, 2008, pp. 137-138). In this later essay he sounds again the "good enough" theme of trying to balance need against trustworthiness. He writes, "[P]lacement is valuable, even necessary on many campuses, but we seem not to have a good way to do it" (p. 136). He finally decides that the flaws are so deep as to trump the need: "Let's face it: almost all of our placement tests are not valid and we shouldn't be using them" (p. 138).

In this chapter, I too am trying to figure out what *good enough* evaluation looks like for various particular situations. The evaluation of writing can never be perfect, but we can try to balance, as Ed did, the need versus the harm or risk.

THREE INHERENT TRAPS OR ILLUSIONS IN THE EVALUATION OF WRITING

1. Trying for a single number score or one-dimensional grade. A single number can never accurately represent the quality or value of a multidimensional entity and writing is inherently multidimensional. Certain dimensions of any piece—for example, the organization, the reasoning, the voice as it relates to the audience, or the spelling—will almost invariably be better or worse than others. No single number will do. Even if one reader thinks that all the dimensions of a piece are of equal value (e.g., B minus), some other reader will weight the dimensions differently.

2. Trying for objectivity. If we accept the premise that writing is for human readers (rather than God or machine-scoring devices), then the *value* of a piece of writing must be tied to the responses of human readers. But humans differ, so different readers will disagree as to value.

Admittedly, human readers often do have single one dimensional reactions or perceptions of a paper (e.g., "This is terrible" or "This is a pure instance of B minus"). But that does not mean we need to settle for naive, global reactions based on holistic feelings ("I like it / I don't like it"). We do not forfeit evaluation by live human readers if we ask them for thoughtful judgment that describes and discriminates between strengths and weaknesses in different dimensions of a piece of writing.

Testers try to escape this second pitfall in various ways:

- They work at "high reliability" in scores by "norming" readers to agree with each other. But in doing this, they simply force

those readers to ignore their own actual differing human responses as to value.
- They enlist similar readers, for example, using only archaeologists for essays meant for that audience. This permits them to announce: "This score represents the value of the writing for archaeologists." But there are still lots of difference between the reactions of different archaeologists.

3. Trying to evaluate someone's skill or ability by looking at a single piece of writing. This trap is most blatant in many placement exams: They give scores to *texts* when the exam is being used to judge ability to write—in this case the *ability* to prosper in one or another first-year writing course.

A moment's thought shows us that the effectiveness of a *single* text or performance can not be a valid picture of a writer's ability. Any evaluation of ability needs to look at multiple performances: texts of various kinds or genres produced on various occasions. And when it comes to evaluating ability, the first two traps also yawn: The ability to write is multidimensional and thus cannot be accurately represented by a single number—and certainly not a number that professes to be objective or fair.

In short, there can be no *single* correct, objective, fair "true score" for any text or any person's ability to write or anyone's likelihood to learn in a given course. This is pretty bad news, and it makes me deeply skeptical of evaluation. I seem to be on the brink of saying what any good postmodern theorist would say: there is no such thing as fairness; let's stop pretending we can have it or even try for it (see Herrenstein-Smith, 1988, on the "contingency of value").

THREADING A PRAGMATIC PATH BETWEEN POSITIVIST FAITH AND POSTMODERN SKEPTICISM

But I am not stepping over that brink. I think fairness is *largely* unavailable, but I would argue that there are situations where it is worth trying to get closer to it on the basis of a pragmatic calculation of *need* versus *danger* (see Baldwin's explorations of fairness, this volume). The main argument of this chapter is that we can figure out the difference between evaluative practices that are *more* fair and *less* fair. There are particular circumstances where the need for a verdict is pressing enough and the danger is reduced enough that it is worth getting a verdict that is only *somewhat* untrustworthy. I am looking for good enough evaluation. ("Good enough" as a positive goal comes

from Winnicott's, 1953, concept of "good enough mother." He was not just suggesting a compromise for tired mommies. He was actively *criticizing* the goal that can understandably tempt new mothers: being the "perfect mother" who fills all the infant's needs. He insisted that "perfect mothers" actually impede growth. In his seminal essay, "Transitional Objects and Transitional Phenomena," he argued that infants actually need a mother who may start off meeting the infant's every need, but who gradually meets fewer and fewer of them to help infants gradually learn more self-sufficiency [p. 1-25]).

I have recently been learning from Hepzibah Roskelly about the rich, broad, stream of philosophic pragmatism that has fed our profession and indeed our country in ways that are too little noticed. The pragmatic approach often involves taking a "third way" that side-steps dead-end conflicts in *theory* and attends to particular cases. So in this chapter, I am trying to side-step the theoretic impasse between a positivistic faith in measurement or assessment and a postmodern skepticism about any possibility of worth in measuring, testing, or scoring.

I am suggesting some general principles in this chaper—bits of theory, yes—but (in the pragmatic tradition) I am refusing to be too rigid in applying those principles as I pick my way gingerly through considerations of particular evaluative situations. The pragmatic move is always to ask *What difference does it make* if you apply a principle in this way in this particular case. Roskelly writes:

> We're never finished with an idea, never completely sure of our conclusions or our directives. For the pragmatists, that's a consequence to be wished for. "We learn to prefer imperfect theories," proto-pragmatist Emerson says, precisely because they're unfinished and capable of change. (personal communication, but see Roskelly & Ronald, 1998)

So I want readers to ask *What difference would it make* if they tried looking at the evaluation of writing through the pragmatic lens I am offering. In what follows I argue that, on the one hand, we need to stop doing some things that are mostly taken for granted. But that, on the other hand, it is possible—and not even so hard—to have many useful evaluations of writing that pass the test of good enough. Putting this point differently: I am trying to play the believing game with evaluation (seeing the needs and possibilities), but also play the doubting game with it (seeing the grave limitations and indeed impossibilities).[1]

THE CLASSROOM AS A LABORATORY FOR A THEORY OF GOOD ENOUGH EVALUATION

I can illustrate my approach by looking at the most pervasive site for the evaluation of writing: the writing classroom itself. In many writing classrooms, teachers put conventional one dimensional grades on individual student papers. Keeping in mind the three traps, it is obvious how deeply flawed such grades are.

- The first pitfall is most obvious here. Conventional grades inevitably mask different teachers' differential weightings of dimensions in multidimensional writing. For example, one teacher might give a B minus to a piece of writing that is very *brilliant* but careless because of confusing organization, quite a few tangled sentences, and lots of surface mistakes. The same teacher might give a C or lower to a paper that is very *careful* (clear, well-organized, and without mistakes), but deeply perfunctory or shallow in thinking. Yet another teacher with different values would give those two papers exactly the opposite grades.
- The second pitfall will also condemn grades if the teacher calls them or implies them to be fair objective evaluations—rather than verdicts deeply influenced by her own values and point of view.
- The third pitfall is usually avoided. Teachers seldom imply that single-paper grades are fair representations of the student's skill. As classroom teachers, it is our stock in trade to say things like, "I know you *can* do better" or "You finally showed your ability to think well on this paper."

It is not surprising that so many students are suspicious and even hostile about the grades they get on their pieces of writing. Almost every citizen of the United States has gotten more grades on writing than on any other school performance in their lives. Understandably, most of these citizens have had experiences that led to resentment and distrust. ("*That was really a good paper but she gave me a C plus on it!*" "*This was a hurried piece of crap where I just told him what he wanted to hear, but he gave me an A.*") I believe that this pervasive and *justified* distrust of invalid teacher grades on writing goes a long way toward explaining why so many citizens and legislators are willing to pay big companies for large-scale exams. Those computerized exam scores (often down to three decimal points) fall into the second trap, of course—pretending objectivity; and they may often test something

different from what we want to test. But the test-makers work harder and get closer to objectivity than rushed and harried individual teachers can manage as they put unilateral grades on papers. This understandably impresses the public.

But there is good news about classroom grading. *Many* classroom teachers have learned to avoid all three pitfalls, and it does not cost them more work: just more care and wisdom. With regard to the first one, they do not settle for a single quantitative score like B minus. They use some kind of grid or rubric or narrative evaluation in order to figure out and communicate what they see as the value of the *various dimensions* of the piece of writing.

Rubrics help these teachers notice and articulate more about a text. Like any reader, teachers often have a global response to a paper and do not know at first which qualities or which dimensions led to this global response For example a teacher might feel, "This paper is very poor. Look at all the surface errors." Yet that teacher actually turn out to feel much more positive about another paper that has just as many surface errors—but the teacher did not notice them so much. Using a grid can help that teacher discover that errors were a red herring in the first case; it was an irritating textual voice or what feels like a noxious point of view that led to the negative reaction. Or perhaps that first paper had "Black errors"—which research has shown to bring down grades more than garden variety "White errors." Rubrics can help readers notice the unfair influence of dimensions they had not been consciously noticing.

Rubrics have come in for some fair criticism when they are crude prepackaged lists of conventional features used on large-scale tests—forcing battalions of readers to try to fit their human responses into corporate pigeon holes. Bob Broad (2003) has written a definitive empirical study of the facts of how individual readers have different responses to the value of writing. But when a rubric fails to include a dimension of the writing that was actually influencing the reader (e.g., voice or point of view), it can tempt a teacher to stay blind to that feature. Almost anything that is *obvious* in writing (e.g., bad reasoning) can be a red herring and mask the influence of other subtler features that actually determined that reader's sense of value.

However a rubric can be used by an *individual teacher*, he or she can design it to fit his or her particular values—and also create different rubrics for different assignments that call for different textual strengths. Or a small group of teachers can collaborate to design rubrics that fit.

Of course, teachers can avoid the problems of rigid rubrics by using only a written comment. I have had lots of experience with this method—especially at Evergreen State College. But I have noticed how I and other teachers sometimes wrote long and thoughtful comments that nevertheless never got around to talking about some crucial and determinative features of the writing.[2]

One of the arguments against rubrics or grids is that they ask for too much work: Teachers have to give five or six grades instead of just one. But teachers who use grids have found a simple solution to this problem: They use only *minimal* verdicts for each item on a grid, namely *strong*, *okay*, and *weak*—or *excellent*, *satisfactory*, *poor*. This means that when they consider each item on a grid, they do not have to *ponder* and try to make careful distinctions. After they read the paper, they hold each criterion in mind for a moment and simply wait to see if a bell goes off in their head saying, *This paper is terrific—or awful with regard to [say] organization*. If the bell goes off, the answer is clear; if not, the answer is also clear: *okay*.

This is not just a lazy short cut. It reflects good evaluative logic for many reasons:

- By giving verdicts on four or five dimensions of a text, a reader is vastly increasing the amount of evaluative information over what we get with a single quantitative verdict or grade. This fits with the theme of *good enough* evaluation. The resulting collection of *crude* grades actually adds up to a richer and more sophisticated evaluation.
- We do well to jettison those hard won attempts to decide between C- and B-level quality on, say, thinking or ideas. They are worthless because readers so often disagree at this level. They enact that perennial hunger for ranking people or performances into fine grade differentiations—when those differentiations are simply not trustworthy.
- The more levels of discrimination of quality are used, the more occasions for disagreement not just by fellow teachers but by students themselves—unnecessary occasions for resenting our verdict and thus undermining the climate for teaching and learning. (I will never forget walking into my office at MIT in the 1970s and finding a paper on the floor that an admirably fearless student had slipped under the door. I had given it a B minus and tried to show in my comments why this was the right grade. But scrawled boldly across the top was this simple message. "This is a B *paper*. Fuck you.").

There's a hoary evaluative principle that says that scorers should never be allowed to be "lazy" and choose a "medium" or middle score. If you believe this, you can avoid three levels and use four instead: poor, fair, good, excellent. But the distinction between fair and good is exactly what we should not trust. And going to four makes for lots more work.

Many classroom teachers avoid the second pitfall too: They have learned not to pretend that their evaluation is objective. They have the courage and wisdom to say something like this when they hand back papers:

> *I cannot pretend that these multidimensional grades are actually fair. Other readers might well give different evaluations. And I want to be clear about something many of you have already come to believe: in fact there is no fair grade—no "true score" for a piece of writing. The best you can hope for is individual readers giving you their most accurate picture of their most careful reading of the strengths and weaknesses of the various dimensions of your paper. That's what I've tried to do. All evaluations will inevitably reflect a readers' own particular values and situatedness.*

Interestingly, when some teachers try to avoid the first trap by using a grid with multiple criteria, this tempts them into the second trap: *I am avoiding the obvious bias that comes with single score holistic grades—grades that are prey to knee jerk global reactions. I will use more concrete objective criteria and that will make me more objective*. But the inherent problem remains: The value of writing is necessarily value for *readers*, and even reactions to particular criteria will differ because they are rooted in the scorer's point of view or cultural situation.

In truth, many teachers find that rubrics help them *avoid* the second pitfall of pretended objectivity. In using a rubric they say, *You deserve to know more about my values as a particular reader: Here are the aspects of writing that I believe are most central to my idea of excellence*. It is particularly helpful for students if we give out the criteria for any given assignment *before* students have to write. It usually results in better essays (or at least essays that suit us better). And we can teach better if we are willing to engage in the self-analysis of figuring out what criteria we care about most—in general and for any particular assignment.

Even when teachers include a literal "bottom line" on their grids—a final line that gives a global *one-dimensional* verdict on the overall quality of the paper—they can still acknowledge their human positionality as readers. When this one-dimensional verdict is part of a grid, it is all the more clear that there is no such thing as a true score. Many sophisticated teachers send a message like this:

> *Here are my perceptions of the quality of the various dimensions of your paper. I've included a bottom line that shows my sense of the overall quality. You can see, thus, how much my global judgment is a product of my personal priorities: how much weight I give to the various dimensions, such as surface features, organization, reasoning, voice &c.*

Let me call attention to the evaluative wisdom in another common practice of many writing teachers: getting students to give each other peer feedback and evaluation. This too tends to avoid the first two pitfalls. Even though peer evaluators are usually less skilled and experienced readers than teacher evaluators, these peer responses are a palpable enactment of a more

valid picture of *value* in writing: They consist of the reactions of multiple and different readers.

RUBRICS, HOLISTIC SCORING, AND CRITERION-BASED EVALUATION

McClelland made the definitive argument against holistic scoring in a 1973 essay in the *American Psychologist.* It was a critique of the tradition of *norm-based* assessment and a proposal to use *criterion-based* assessment instead (or outcomes-based or competence-based or mastery-based assessment). The problem with norm-based evaluation is that it gives us nothing but a number: no information about what the student actually knows or can do. *B minus* or *85* tells us nothing about what students have learned or what they can do. So holistic scoring is a norm-based enterprise. Admittedly, in large-scale assessments, administrators try to diminish this problem by writing a "guide" that is supposed to tell what a "4" or "2" essay looks like. But these descriptions are notoriously unsatisfactory as portraits of the actual essays; they offer a kind of Platonic picture of the ideal mix of features in any given score.

The goal for norm-based evaluation is an ideal that is seldom realized: A set of scores that fall into the pattern of a bell shaped curve: the maximum distribution among skills or abilities (or intelligence). The goal for criterion-based evaluation is a list of things students should have learned—and for each item a *yes* or *no*.

This insistence on a binary yes/no result for each outcome is a problem that bedeviled the outcomes-based or competence-based movement and helped lead to its fading. Too many things that we teach and want to evaluate are not susceptible to black/white yes/no answers. The problem is particularly obvious with writing. For example, is a given essay competently organized—or well thought through—or well adapted to its audience? For some essays we can give a clear *yes* and/or *no* on each criterion, but many essays force us to answer, *partly* or i*n some ways yes but in other ways no*. That is, the criterion-based folks were obviously right to insist that large multidimensional entities like a text—or abilities like *writing*—should be broken into smaller pieces. But it is hard to break them down so far that evaluation results can consist of pure *yes*'s and *no*'s.

Rubrics come to the rescue with this problem. Rubrics represent a move away from *norm-based* evaluation (or holistic scoring—using only numbers) in the direction of a *criterion-based* process that insists on articulating what is to be learned. Yet when rubric users insist on scores of 1 to 5 on each

criterion, they fall back into the norm-based trap: fixating on fine numerical distinctions that will not hold up. But when we use rubrics with only three levels of accomplishment, for example, *strong*, *okay*, *weak*, we are not just settling for a *compromise* between norm-based and criterion-based evaluation (not that there is anything wrong with compromise). I would call it a *good enough* approach that is actually better than either alternative. The cruder scores are much easier to give and the results are more trustworthy.[3]

BUT CAN ONE DIMENSIONAL OR SINGLE NUMBER GRADES BE GOOD ENOUGH?

I haveve talked of traps, pitfalls, and failures but have been at pains to show how thoughtful teachers manage to avoid those traps and still give quantitative classroom evaluations to individual papers and portfolios. But I have stressed that those evaluations are only good enough because they do not consist of single-number grades. Now, however, I want to look at situations where I think single-number grades—one-dimensional verdicts—can manage to be good enough.

For a writing prize. Teacher are sometimes asked to nominate a student for a writing prize. A nomination would seem to fall right into the first two pitfalls: It involves a misleadingly single number (a yes/no decision) based on a biased judgment. And the consequence can be weighty—sometimes significant money. Nevertheless, I would defend such a nomination as a good example of "good enough" evaluation. I show how such a nomination relates to the three pitfalls.

With regard to the first pitfall, a single-number score does much less harm at the extremes of quality—excellent or poor. That is, the biggest unfairness in single-number grades comes from the way different evaluators disagree in the weight they give to different dimensions. But when a single reader calls a paper or portfolio *excellent*, those differential weightings are a little less likely to do harm. Excellent features are more likely to predominate—or else one particular feature may be so strong as to overshadow other weaknesses—even in the reactions of others readers. Therefore a somewhat larger proportion of readers is more likely to agree that the paper or portfolio is excellent (or poor) than will agree about a single-number grade in the middle range of B or C where differential weightings more easily tip the balance. Let me emphasize: I am not saying that one teacher's "outstanding" will garner agreement from all readers; but at least more of them are more likely to call it notably good than will agree about some middling grade

where the mix of dimensions is killing (see Despain % Hilgers, 1992, for some research backing up the idea that decisions at the margins are a bit more reliable).

And when it comes to the second trap—objectivity versus bias—the danger is even smaller. For in almost all prize situations, the teacher is not *awarding* the prize—only nominating a student. There Is a committee that must adjudicate. In fact, the awarding of prizes for excellence in writing reflects a remarkably valid and sophisticated understanding of how the evaluation of writing ought to work. The prize is given as a result of negotiation among necessarily biased evaluations by situated readers. And when it comes to writing prizes, almost none of the participants or audiences has any illusion that they are looking at a "true score." WhereAS people tend to read grades as professing to be "valid," most observers see prizes as contests of *taste*. So they can see the process is an *attempt* at a certain kind of fairness—with full open recognition of the impossibility of attaining it.

If it is a prize for a body of work, the third pitfall is avoided, because it is based on multiple texts. But even if it is a prize for just *one* essay or story or poem, there is very little pretense that the prize is actually a measure of the writer's true ability. Everyone can see that it is a prize for one performance—and that might well be better or worse than the writer usually manages. It is like a gold medal for the hundred yard dash: People know that the winner is sometimes not as consistently good or skilled a runner as someone else who happened to have had a bad day or even bad season.

Failing a student for the course. Here is another one-dimensional score, yet there is even more pressure for fairness because the consequences are so weighty: no credit and the requirement to take the course again. I can continue to clarify my theory by arguing how a failing grade can make evaluative sense as "good enough"—*but* with one important reservation. Let's look at the calculus of need versus danger.

- The need is great. For teaching and learning to go on in institutions that give credit, it is important and valid to be able to withhold credit and require students to learn enough before they go on to future courses.
- The danger is not so very great. The single-number verdict is not so damaging because it is at the extreme. When the teacher decides she should fail a student because of poor writing, the writing will surely be very poor and there will be significantly more agreement among readers (of course, not complete agreement). With very poor writing, the disagreement will be less than about, say, the grade of C or C plus.

But, of course, we cannot *fully* trust a unilateral judgment to be fair, even at the margins. And when it comes to failing students, the *feelings* of teachers tend to play a big role and feelings are notoriously unfair. (*I just can't fail this student who's been so diligent—in fact he was a big help to me in teaching this class." "This student has been a complete pain the ass. I'm glad her writing looks so awful to me." "I can't fail someone who's been taking care of his dying mother."*) So I would argue that we must not accept such a weighty consequence as a failing grade if it is based on just *one* person's judgment about the quality of writing. It is not "good enough" evaluation to fail a student for a course unless at least one other instructor shares in the decision.

This is not so hard to pull off. It is more or less what Pat Belanoff and I set up at Stony Brook. If a teacher wanted to fail a portfolio and thus deny credit and require retaking the course, another teacher—who did not know the student—had to agree that the portfolio was of failing quality. (In fact, in my experience as a WPA at two universities, few failing grades were based *mainly* on quality of writing. Usually they stem from some dereliction of duty—and for deciding on that, there is no need for a second opinion. This chapter is about the evaluation of writing.)

Eligibility to keep a scholarship or play on a varsity team. "Professor, you just have to give me a B or I'll lose my scholarship [or be kicked off our winning basketball team]." Teachers are often asked to sign forms with a single number "score"—that is, to certify that students on a team have a B or B minus average in the course. Here it seems clear to me that such evaluative decisions *do not* make sense—they are *not* good enough (unless the student's performance is *massively* excellent or poor). They fall squarely into the first two traps. They represent single-number grades for multidimensional performances that fall right in the middle of the scale where disagreement among evaluators is virtually inevitable—and they have to pretend to be objective or fair. Think back to those two teachers I spoke of earlier who were dealing with brilliance and carelessness in two matching papers. The very same student would have kept his scholarship if he had had one teacher but would have lost it if he had had the other one. If a grade determines an important consequence like keeping a scholarship or being on a team, we need fairness.

This conclusion may cause problems: *We need some way to decide on eligibility for keeping a scholarship or being on a team!* But there is no need to decide on the basis of fine-grained decisions about quality of writing in the middle range. We can be open about other criteria that probably play a bigger role in such teacher decisions anyway, for example, how well students are meeting all the concrete obligations of the course (such as attending, getting assignments in on time, doing substantive revisions, and so on).

What course grade should the student get on the transcript? I have tried to justify a final course grade of F, but what about all the other final grades? One good thing about them is that they are almost always based on *multiple* and *different kinds* of writing, not just one text (although there are some upper level and graduate courses where teachers base the final grade on just one big term paper or exam). But course grades fall squarely into the first two pitfalls: They are single global numbers meant to represent the value of multidimensional performances, and they are meant to be fair when in fact they are unilateral judgments by just one reader with one inevitably partial point of view and set of values. (I have defended nominations for a prize because that unilateral judgment is simply a doorway into a collaborative judgment. I have insisted that failing grades are not good enough unless they are collaborative.)

This problem is all the more serious because the stakes are high. Compared to the grade on a single piece of writing, the course grade goes on the transcript, and there are many readers who use it to make weighty decisions. It affects grade point average and graduation, applications to graduate schools and employers. Yet the same student would likely get different final grades from different teachers looking at the same work—at least with mid-level quality work. A course grade of *B minus* will actually mean many different things that readers have no way of fathoming: It could mean *pretty good in all aspects of the course*; it could mean *brilliant writing but great carelessness and irresponsibility in meeting responsibilities*; or it could mean *rather poor skill in writing but lots of growth and enormous diligence in all other respects*.[4]

Technically speaking, it would be easy for institutions to stop giving those untrustworthy conventional course grades, but few have had the wisdom or taken the trouble. It would simply require that course grades come in the form of a grid through which the teacher can communicate more clearly how well students have attained the skills or abilities or understandings that are asked for in the course. Grade readers never get this information from a conventional transcript, yet it is just what most of them need for making the decisions that they normally make when they consult a transcript.

Here are some things that grade readers typically want to know when they read course grades for writing courses: *How well can students think and argue on paper? How clearly can they make their points and their sentences? How skilled are they at mastering the conventions? How diligent and responsible were they in meeting obligations?* (Readers of grades in other courses tend to have other questions: *How well have students mastered the concepts? How well can they apply the concepts to new material? How clearly can they write about the course material. How diligent and responsible were they in meeting obligations?*)[5]

Multidimensional final grades would require a bit more thinking from teachers—helpfully asking them to be more self-aware about what skills or abilities they are trying to teach. But they would not actually require much more grading work, because again, *minimal crude* verdicts would be fine on each criterion: *strong*, *okay*, *weak*. I think most teachers would be relieved to turn in grades that were more accurate and less misleading.

There is no need for all teachers to agree on one set of criteria for course grids. Indeed, teachers *should* make their own decisions about what dimensions of performance are most important for their course. Transcripts with multidimensional grades could be handled easily by the registrar with computers, and they would be far more accurate, fair, and useful as evaluations of student learning. Why else did God invent computers if she did not want us to communicate learning more specifically and clearly?

Should these grid grades contain a "bottom line" single-number holistic grade? Why not? A global grade is not so opaque when accompanied by the other items in a grid. Readers can read those global verdicts more critically and usefully. They can tell, for example, that a student got a good course grade even with a "weak" for memory—or a poor grade for the course even with a "strong" for diligence and responsibility.

I do not know any college that uses a grid transcript, but it is exactly what report cards look like for children in the early elementary grades. Most teachers and policy makers seem to feel it is "childish" to have multidimensional grades—when in fact it is much more sophisticated and evaluatively valid. Evergreen State College and Hampshire College and a few other places use narrative evaluations on their transcripts instead of single-number grades.

But what about bias and fairness? Even though these sophisticatedly multidimensional grades are more informative, they would still be unilateral judgments made by individual teachers with limited and inevitably biased points of view. The registrar would have to print a disclaimer on each transcript: "The college makes no claim of fairness for any of these grades."

Yet interestingly, I do not see such a big problem here. For in fact, most readers of transcripts read course grades to mean something like this: *This grade for this course represents what this teacher thought of the student's performance, while that grade for that course represents what that teacher thought of her performance*. There's something salutary about the genre of a transcript—especially in its visual form. The sight of all those individual teachers' grades crowded together next to each other tends to disabuse people of any illusions that they are seeing "true scores." (Perhaps I am not acknowledging naive readers who feel that a transcript is a perfect X-ray of the student's entire ability. I think there were more people who fell into that assumption 50 years ago than in our rather cynical age.)

But GPAs? They fall deeply into the first two traps. They would be based on those frail bottom line scores on teacher grids. They would be outrageous failures to represent the *myriad* dimensions of a student's learning and performance. And once you reduce the multifarious complexity of a transcript to a single number, an unwarranted implication of objectivity or fairness sneaks back in. I would argue against computing a GPA for any single semester or year's performance. However, there might be a kind of rough quantitative GPA—both good enough and useful—it is simple to compute how many bottom line As and Fs a student got as a percentage of all grades. (Note how this discounts all the middle-level untrustworthy grades.)

APPLYING THE THEORY MORE WIDELY

If we apply this theory of good enough evaluation more widely, we see a combination of good news and bad. It tells us to give up certain convenient practices and handy scores; but in many cases we can compensate. The payoff is more trustworthy evaluation—and evaluation that will give rise to much less cynicism.

Placement exams. In the 1980s, there were more placement exams in the United States than any other kind of writing exam. (This is the finding of three research reports: CCCC Committee on Assessment, 1988); Greenberg, Wiener, & Donovan, 1986); Lederman, Ryzewic, & Ribaudo, 1983; cited in Greenberg, 1992, p. 17.) White says that conventional holistic scoring of placement essays is good enough. I disagree. But again, my goal is the same as his: Not a perfectly trustworthy evaluation, but one that is good enough.

When we run the calculus of need against harm, need loses out. These thousands and thousands of placement tests are largely unnecessary—and I call them harmful. There is a remarkable array of practices and writings that show us good alternatives to placement testing. The most elegant and easy one is Directed Self Placement (see Royer & Gilles, 2003). It is no longer a new and odd experiment; it has been used with satisfactory results in a very wide range of institutions.

And there is another alternative model (equally widely tested and used) that is also better than conventional placement tests. Students who need more help to prosper in the regular first-year course are identified, but not in a big test. Instead the identification is made by the regular teacher in the regular course classroom in the first week. Such students are then obliged to attend a one-credit course—or in some places a workshop—and this functions as a *supplement* to the regular first-year writing course. Students so placed get lots

of extra help of all kinds so that they can stay and learn in the regular course—avoiding the ghetto effect of segregating them so they never get to work with stronger more confident students (among other sources, see Benesch, 1988; Elbow, 1996; Grego & Thompson, 1995; Kidda, Turner, & Parker, 1993).[6]

Evaluation of programs. In the last decade or two—especially with No Child Left Behind—the amount of placement testing has surely been exceeded by *programmatic* evaluation of writing.

> *Have we succeeded in improving the writing of the students in our school district [or high school or middle school or first-year writing course or lower division program or entire college curriculum]? Can we demonstrate adequate yearly progress?*

Let's look at the three pitfalls.

1. Like placement tests, these program evaluations typically use holistic scores—single numbers that fail to represent the value of a multidimensional product.
2. They typically pretend to be fair and objective.
3. They typically pretend to measure *change in ability* by looking at only two texts—"before" and "after" essays that testers try to make as absolutely similar as possible. Students often have a better day for the "before" essay than the "after" essay.

This looks pretty bad, but again let me try to show the possibility of a good enough alternative. I think the biggest help will come from avoiding the third pitfall. Plenty of programs do this by using *portfolios* for the before and after snapshots—even small portfolios. Thus they are looking at multiple texts in multiple genres that were produced on different days. It is better still when programs get papers in the portfolios that represent what most of us would call "real writing"—that is, writing where students had a chance to draft, get feedback, and revise. After all, the ability to write under exam conditions with no chance to revise is, surely, not what most people mean by "writing ability." Of course, it is expensive to use portfolios for programmatic evaluation, but there is no way around it if we want good enough results. But it makes perfectly good sense when institutions deal with high costs by *sampling* students instead of trying to evaluate them all.

Is there any way to avoid the first two traps: using single-number scores that pretend to represent fairly the value of multidimensional pieces of writing? It would seem impossible. Program evaluations tend to use not just single-number scores but single-number *averages* of single-number scores. And they usually have to talk about very small quantitative improvements.

Hooray, we've moved the average from 2.8 to 3.2. We are a success!
Oh dear. 2.95. We're a failure.

I have seen it happen. It is *exactly* these kinds of small single-number scores in the middle range—the range where most students live and where most readers disagree—that are least trustworthy.

But there is a good enough route around this first trap. Let's think about the goal of programmatic evaluation: to see how well a program is doing and try to improve it—that is, improve teaching and learning. Most legislation that requires assessment requires thoughtful examination of what is working well and not so well. This goal *can* be well served without single-number scores. Think back to the virtue of grids. What could be better than having readers score *multiple* criteria as they read portfolios? This need not be a killing task because, again, readers can rank criteria on only three levels—weak, okay, and strong. This is not just cost cutting; in fact, it makes the results more trustworthy.

Here are some abilities that might be scored: The ability to mount an effective train of thinking; to support it with evidence and examples; to demonstrate a sense of audience and genre; to create a structure or organization that is effective for most readers; to manage conventions; to write about personal reactions and feelings tellingly—as a valuable skill in itself, but also as part of a less personal argument. Some programs might want to evaluate more specialized and particular criteria of educational growth. For example, a general education curriculum—whether for the lower division courses or the whole college—might want to evaluate how well student writing demonstrates some understanding of cultures different from their own. (Students need to know in advance, of course, that they need to put together a portfolio in which some of their papers show this kind of thing.)

There is usually no requirement that a program try for a single number to represent improvement in *overall* writing ability. If we ask readers to rate different *dimensions* of writing skill in a portfolio, there will be some program-wide differences that are strong enough to be meaningful and useful. Certain dimensions of writing skill will show more improvement—or less. Even portfolios that are middling as a whole will often show substantive strengths and weaknesses on certain dimensions. This approach has a chance of revealing at least a few meaningful numbers that could usefully guide curriculum planning, course planning, and teaching. How much more meaningful and useful to all "stakeholders" to tally the number of *strongs* and *weaks* for each criterion. Thus one might be able to conclude, "*Between September and June, with respect to 'mounting an effective train of thinking,' there were twenty percent fewer weaks and fifteen percent more strongs.*" It will sometimes be possible to see which dimensions showed more improvement and which ones less.

I seem to be talking as though single-number scores can never be useful. But there are exceptions. A few portfolios will be globally and strikingly strong—or weak—in *most* dimensions. I invoke here my earlier justification for nominating a student for a prize or giving a failing course grade. A one-dimensional score or verdict—although not wholly trustworthy—can be good enough to be used when it is at the extreme. That is, I would argue that programmatic evaluations could validly identify writers whose before and after portfolios show their degree of improvement near the top of what can be expected—and also those whose degree of nonimprovement puts them at the bottom. These more trustworthy single numbers would be suggestive and useful, even though they speak of only a minority of students.

Finally, I can quickly describe a way of avoiding the second trap in program assessment: not pretending fairness or objectivity in measuring performances against a stable, objective, universal skill in writing. A program can avoid this pretense with an honest adjustment to the goal of the enterprise:

> *We are not pretending to a measure of some universal Platonic skill in writing. We are only trying to measure student improvement in the kinds of writing we care about at this institution. In short, our target is frankly biased, but it's the target that matters to us.*

A VISION FOR SAT ESSAY TESTS, NAEP ESSAY TESTS, STATEWIDE ESSAY TESTS MANDATED BY NO CHILD LEFT BEHIND, GENERAL EDUCATION ESSAY TESTS, AND ESSAY TESTS FOR LICENSING TEACHERS

The simplest, cheapest course is simply to scrap these tests. They tend to be used for high-stakes decisions with big consequences: A "score" that counts hugely for college admission, graduation from high school or college, eligibility for the next grade or for upper division status, or getting a license to teach. Yet the scores are deeply untrustworthy. The tests usually look at single texts and score them with one-dimensional single scores that are alleged to represent the real value of the text, and the ability of the writer. Falling so deeply into all three traps, their scores are not even useful at the extremes. Surely we have better things to do with money than give tests that give worthless results and create so much unhelpful anxiety.

But I will end the chapter with a vision of how even these large-scale exams that look only at single texts *could* be far more useful and valid—and far less damaging. We have to radically adjust our sense of the goal for these exams. They are no good for making high-stakes decisions; but they could

be useful for learning and teaching. I see a large-scale test—district-wide, state-wide, nation-wide—where students can submit a paper they have revised. Each paper would be read and evaluated by three readers, but they would use multidimensional grids. The test administrators would assemble groups of good but representative and different readers. There would be no pretense at "training" or "calibrating" them to make them ignore their own values. Instead they would be invited to read like the human teachers that they are—in all their diversity. Scores would consist of verdicts on, say, four or five rubrics and come from three representative readers. Naturally these results could not be used for any important decisions. You could not rank students or states or classrooms or districts. Yet these diverse reader evaluations would be highly useful to the students and to their teachers. And they would be enormously interesting too for people interested in evaluation.

CONCLUDING THOUGHT

I have been trying to show here the courage of my pragmatism. Some "coherent" thinkers will say I have been far too purist—disallowing too much useful evaluation. Others will say I have been far too permissive—condoning too much dubious evaluation and acceping even *scores* that are so far from fully valid. But even though there *is* no accurate or true or fair score for any piece of writing—and no piece of writing can give a valid picture of a person's skill or ability to write. Nevertheless, we do not need to throw up our hands and reject all evaluation. Like White, we can try to use the calculus of need versus harm. Are there conditions where we need some kind of judgment strongly enough and where the danger of untrustworthy results is reduced enough that it is worth going forward with the evaluation? People and institutions need to make this calculation for their circumstances; there is no one right way to evaluate writing for everyone and in every context. In dealing with this puzzle, it is salutary to remember the wide range of evaluations that go on in the world. Consider the processes of hiring someone for a job or accepting an article for publication. The process is usually like the one used in awarding a prize. Usually it represents a negotiation of multiple perspectives.

Nevertheless, I have tried in this dchapter to use this kind of calculus for many educational settings. I end up working my way to deciding that the following evaluations are worth making if the stakeholders want them:

- Individual teachers giving multidimensional feedback or scores on individual papers (as with grids or rubrics). But only three levels of quality are warranted for each dimension.

- Individual teachers nominating students for a writing prize.
- Individual teachers giving a failing grade for a course (a one-dimensional verdict) on the basis of poor writing—if another teacher concurs.
- Individual teachers giving course grades for a transcript—as long as those grades are multidimensional.
- For 2 or 4 years of work, computing not a GPA, but a minimal cumulative "score" consisting of the number of bottom line *strongs* and *weaks* compared to the number of all courses taken.
- Using portfolios for programmatic evaluation to identify improvement or lack of improvement on various dimensions of writing ability.

If my calculus for good enough evaluation seems too austere and disqualifies too much evaluation that people think is necessary, let's not forget a different calculus of need versus harm. The need is for money: Evaluation is very expensive, and we need more money for teaching and smaller class sizes. This calculus makes it all the harder to justify many of our current evaluations that are so expensive, dubious, and that so often do great *harm* to the climate for teaching and learning.

ENDNOTES

1. I look at evaluation through an eclectic but extensive set of experiences and writings since 1969. I taught at Evergreen State College for 9 years—where we used no grades at all. I spent 4 years as part of a research team looking at a dozen experiments in competence-based higher education. Pat Belanoff and I started the movement for using portfolios for program-wide evaluation (though I have since found a tiny Hawai'ian religious college that beat us to it). I have published more than 20 essays about assessment (for a list of my works cited, see http://works.bepress.com/peter_elbow).
2. See Elbow (1997) on this issue. It includes an appendix of the many publications arguing against holistic scoring.
3. There was a big movement in the 1970s for "competence-based" evaluation. I spent 4 years on an 8-person research team investigating competence-based programs in higher education. (We each had a site but we all visited all the sites and each wrote field notes on our visits. See Grant et al., 1997, and Elbow, 1979. Remarkably, David Riesman was one of our team.

   For a long time it seemed as though the enthusiasm for competence- or outcomes-based education had faded away. Perhaps the approach asked for too much from teachers. It required teachers to figure out specifically what they want to students to learn or be able to do—and also to articulate these learning

goals publicly and clearly enough for students to understand them. *And* in addition, of course, to figure out a way to evaluate whether the students have learned or can do those things. Also, the problem of asking for unambiguous yes/no answers tempted practitioners into smaller and smaller outcomes—sometimes to the extent of tiny *behavioral* objectives (*Are there paragraph breaks at least every 10 or 12 sentences?*) Also competence-based enthusiasts sometimes betrayed a rhetorically unhelpful resentment against conventional college professors who said, in effect, "Don't ask me to specify exactly what I'm trying to teach. Only *I*—the expert in this area—can say what it is, and you wouldn't understand."

In the last decade or two, however, we have seen a resurgence of the criterion-based spirit with the growth of interest in *outcomes*—across all fields from business to government to education. Note the Outcomes Statement approved by so many members in the association of Writing Program Administrators (2000, 2008). Many of the outcome bulldozers lead to crude and unhelpful results, but I cannot help thinking that the essential wisdom in the criterion-based impulse sparked by McClelland cannot be kept down. It gradually dawns on more and more people that it is useless and harmful to evaluate unless the results involve *words for describing what is being evaluated*—instead of just numbers for ranking people as better than or worse than.

4. The method of contract grading that Danielewicz and I (2009) wrote about in CCC avoids all the untrustworthy mid-range grading and asks us to use a single-number score only for outstanding performances—and also to wait until there are multiple texts to base it on. Yet that scoring does not get in the way of the useful evaluative feedback we give to individual papers. Because the contract focuses on behavior, it lets teachers spend very little time trying to evaluate behavior. Instead they put their time and energy into giving writerly feedback and figuring out which student behaviors to require—that is, which behaviors most reliably lead to learning to write better.
5. What about teachers of large lecture courses in science who base course grades on nothing but one or two machine-graded exams? They can still usefully give a course grade of more than one dimension. When they make up such exams, they are usually conscious (or need to be) or whether a question tests memory, or a theoretic understanding of concepts, or an application of concepts to new material, or computational skill.
6. The evaluative harm from conventional holistically scored placement testing is obvious enough: It falls into all three traps. Most striking is the third trap of using a single text (written under the worst of exam conditions) to judge a student's *ability* to thrive in the regular course. In truth, a test of students' ability to handle alcohol would probably be a more valid measure of how they will fare in first-year writing. The first trap is also lethal: using single-number verdicts for multidimensional entities. With regard to the second trap, conventional holistic scoring on placement tests usually works hard to avoid bias—using two readers and a third in cases of wide divergence. But that process—and the "norming" of readers that goes along with it—just shows susceptibility to the myth of a "true score." See Smith (1993) on a shrewd attempt to avoid that problem: using readers from the courses themselves. These readers are not trying for true scores,

they are asking frankly positional questions of each text: "Does this writing look to me like it was produced by someone who could learn and prosper in my regular section of first-year writing?"

REFERENCES

Benesch, S. (1988). *Ending remediation: Linking ESL and content in higher education*. Washington, DC: TESOL.

Broad, B. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring* (pp. 263-276). Portsmouth, NH: Boynton/Cook-Heinemann.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.

CCCC Committee on Assessment. (1988, Spring). *Post-secondary writing assessment: An update on practices and procedures*. Report to the Executive Committee of the Conference on College Composition and Communication.

Council of Writing Program Administrators. (2000, 2008). *Outcomes statement for first-year composition*. Web.

Despain, L., & Hilgers, T. L. (1992). Readers' responses to the rating of non-uniform portfolios: Are there limits on portfolios' utility? *WPA: Writing Program Administration, 16*(1-2), 24-37.

Elbow, P. (1979). Trying to teach while thinking about the end: Teaching in a competence-based curriculum. In G. Grant et al. (Eds.), *On competence: A critical analysis of competence-based reforms in higher education* (pp. 95-137). San Francisco, CA: Jossey-Bass.

Elbow, P. (1997). Writing assessment: Do it better, do it less. In W. Lutz, E. White, & S. Kamusikiri (Eds.), *The politics and practices of assessment in writing* (pp. 120-134). New York, NY: Modern Language Association.

Elbow, P. (1996). Writing assessment in the twenty-first century: A utopian view. In L. Bloon, D. Daiker, & E. White (Eds.), *Composition in the 21st century: Crisis and change* (pp. 83-100). Carbondale, IL: Southern Illinois University Press.

Elbow, P., & Danielewicz, J. (2009). A unilateral grading contract to improve learning and teaching. *College Composition and Communication, 61*(2), 244-268.

Grant, G., Elbow, P., Ewens, T., Gamson, Z., Kohli, W., Neumann, W., Olesen, V., & Riesman, D. (Eds.). (1979). *On competence: A critical analysis of competence-based reforms in higher education.* San Francisco, CA: Jossey-Bass.

Greenberg, K. (1992). Validity and reliability: Issues in the direct assessment of writing. *WPA: Writing Program Administration, 16*(1-2), 7-22.

Greenberg, K., Wiener, H., & Donovan, R. (1986). Preface. In K. Greenberg, H. Wiener, & R. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. xi-xvii). New York, NY: Longman.

Grego, R., & Thompson, N. (1995). The writing studio program: Reconfiguring basic writing/freshman composition. *WPA: Journal of Writing Program Administrators, 19*(1-2), 66-79.

Grego, R., & Thompson, N. (1996). Repositioning remediation: Renegotiating composition's work in the academy. *College Composition and Communication, 47*(1), 62-84.

Herrenstein Smith, B. H. (1988). *Contingencies of value: Alternative perspectives for critical theory*. Cambridge, MA: Harvard University Press.

Kidda, M., Turner, J., & Parker, F.E. (1993). There is an alternative to remedial education. *Metropolitan Universities, 3*(3), 16-25.

Lederman, M. J., Ryzewic, S., & Ribaudo, M. (1983). *Assessment and improvement of the academic skills of entering freshmen: A national survey*. New York, NY: CUNY Instructional Resource Center.

McClelland, D. C. (1973). Testing for competence rather than for intelligence. *American Psychologist, 28*, 1-14.

Myers, M., & David Pearson, P. (1996). Performance assessment and the literacy unit of the new standards project. *Assessing Writing, 3*(1), 5-29.

Royer, D. J., & Gilles, R. (Eds.). (2003). *Directed self-placement: Principles and practices*. Cresskill, NJ: Hampton Press.

Roskelly, H., & Ronald, K. (1998). *Reason to believe: Romanticism, pragmatism, and the possibility of teaching*. Albany, NY: SUNY Press.

Smith, W. (1993). Assessing the adequacy of holistically scoring essays as a writing placement technique. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Cresskill, NJ: Hampton Press.

Winnicott, D. W. (1953). Transitional objects and transitional phenomena—A study of the first not-me. *International Journal of Psycho-Analysis, 34*(1), 89-97.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication, 46*(1), 30-45.

White, E. M. (2008). Testing in and testing out. *Writing Program Administration, 32*(1), 129-142.