2010

Good Enough Evaluation

Peter Elbow
University of Massachusetts - Amherst, elbow@english.umass.edu

DRAFT ESSAY FOR *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White*

Peter Elbow 2/10

Good Enough Evaluation : An Overall View of When and How Evaluation Makes Sense

Ed White's approach to assessment has always been practical and realistic. Over and over he's 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's what I'm trying to do here.

Famously, Ed argued that holistic scoring on placement tests, though quick and dirty, is good enough for the purpose. I'll describe below why I disagree with this particular conclusion, but I think I'm carrying on Ed's nonpurist approach. I'm trying to figure out what is *good enough* evaluation in various particular situations. The evaluation of writing can never be perfect, but we can try to calculate--as Ed did--when the *need* is great enough and the *harm* or *risk* of invalid results is small enough to warrant going ahead with some particular assessment.

Three Inherent Traps or Illusions in the Evaluation of Writing

There are three pitfalls in the evaluation of writing.

(1) Trying for a single number score or grade. A single number can never accurately represent the quality or value of multidimensional entity like a piece of writing. Writing is inherently multidimensional, and certain dimensions of the any piece--for example the organization, the reasoning, the voice as it relates to the audience, or the spelling--will almost always be better or worse than others. Thus a single number cannot represent these differing values or the relationships among them--relationships that different readers will weigh differently in reaching a single number score.

(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 response of the audience: how well does the text works for its human readers? But humans differ, so different readers will have different responses--among other things as to its effectiveness or value or excellence. Thus there can be no *single* correct, objective, fair measure of the value of a piece of writing. There is no "true value" for writing.

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 just because assessment must be based on the value that live human readers see in a text, that doesn't mean we need to settle for naive, global reactions based on holistic feelings ("I like it / I don't like it"). We don't forefeit 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 a skill or ability by looking at a single piece of writing. Many exams are do what placement exams do--sometimes unawarely: they give scores to texts when the exam is being used to judge the *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 dangers also apply: 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 is no single true score for any text or any person's ability to write. 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 on the "contingency of value").

But I'm not stepping over that brink. I think fairness is *largely* unavailable, but I'd argue that there are situations where it's worth trying to get closer to it on the basis of a pragmatic calculation of *need* vs. *danger*. The main argument of this essay 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's worth getting a verdict that's only *somewhat* untrustworthy. I'm looking for good enough evaluation. ("Good enough" as a positive goal comes from Winnicott's concept of "good enough mother." He wasn't just suggesting a compromise for tired mommies. He was actively *criticizing* the goal of being the "perfect mother" who fills all the infant's needs. He insisted that this actually retards growth. He argued that how infants actually *benefit* from having a mother who may "start off with an almost complete adaptation to her infant's needs," but "as time proceeds she adapts less and less completely, gradually, according to the infant's growing ability to deal with her failure" ["Transitional Objects"].)

I've recently been learning from Hepzibah Roskelly (1998) 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 particular cases. I'm trying to side step the theoretic impasse between a positivistic faith in measurement and assessment and and post modern skepticism about any possibility of value in measurement, testing, or scores.

I am suggesting some general principles in this essay--bits of theory, yes--but I am refusing to be too rigid about those principles as I use them to pick my way gingerly to conclusions based on particular situations for evaluating writing. The pragmatic move is always to ask

"What difference does it make?" if you apply a principle 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. (This from an email, but see her *Reason to Believe*.)

I want readers to ask, "What difference would it make?" if they try looking at the evaluation of writing through the pragmatic lens I am offering. So in what follows I'll argue that we need to stop doing some things that are taken for granted. But I'll also argue that it's possible--and not even so hard--to have many useful evaluations of writing that pass the test of good enough. Here's a different frame for my approach here: I'm trying to do justice to what I learn by playing the believing game with evaluation (seeing the needs and possibilities); but also by playing the doubting game (seeing the grave limitations and indeed impossibilities).

FOOTNOTE. The conclusions I come up with are not pretty or neat. I fear they will look merely idiosyncratic, subjective, or amateur. I can't resist saying what some readers will not know: In fact I have been thinking and writing about evaluation and assessment since 1969. (I have taken the liberty of listing, after my works cited, twenty essays about assessment that I have published.) I taught at Evergreen Stage College for nine years--where we used no grades at all. I spent four 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 use portfolios for program wide evaluation (I've since found a tiny Hawaiian religious college that did it first.) From this vantage point, I see a way to pull a lot of thinking together into an admittedly untidy overview train of thinking.

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 pitfalls, we can see what many people already see, namely, how deeply flawed such grades are.

The first pitfall is most obvious here. Conventional single number grades cannot fairly represent the quality multidimensional pieces of writing. Conventional grades inevitably mask different teachers' differential weightings. For example, one teacher might give a B minus to a piece of writing that is brilliant but careless: it's poorly organized and has quite a few tangled sentences and lots of surface mistakes. The same teacher might give a C or lower to a paper that's 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 that teachers call or imply to be fair objective evaluations rather than verdicts deeply influenced by the values and point of view of the grader. As for the third pitfall of implying that these grades on single papers are fair representations of the student's skill: teachers don't so often make this mistake. As classroom teachers, it's our stock in trade to say things like, "I know you can do better" or "You finally showed your good thinking on this paper."

It's 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 U.S. has gotten more grades on pieces of 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 in fact 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 impresses the public.

But there's good news about classroom grading. *Many* classroom teachers have learned to avoid all three pitfalls, and it doesn't cost them more work: just more care and wisdom. With regard to the first one, they don't settle for a single quantitative scores like B minus. They use a narrative evaluation or a grid of some sort 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 don't 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 how all the surface errors." Yet that teacher might turn out to feel much more positive about another paper with just as many surface errors. Using a grid can help that teacher notice that errors were a red herring; it was an irritating textual voice or what feels like a noxious point of view that led to the negative reaction. Indeed, that first paper might have had "Black errors"--which research has shown to bring down grades more than garden variety "white errors." Rubrics can help readers notice the influence of dimensions they hadn't been consciously noticing.

Rubrics have come in for some fair criticism when they are crude prepackaged lists of conventional features that are used on large scale tests--forcing force battalions of readers to try to fit their human responses into corporate pigeon holes. Bob Broad has written a definitive empirical study of the facts of how individual readers have different responses to the value of writing (*What We Value*). When a rubric fails to include a dimension of the writing that was actually influencing the reader (for example, voice or point of view), it can tempt a teacher to stay blind to that feature. Almost anything that's *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. But when a rubric is 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. Teachers can avoid the problems of rigid rubrics by using only a written comment, however I frequently notice how I and other teachers write long and thoughtful comments that nevertheless never get around to talking about some crucial and determinative features of the writing. (See my "Do It Less, Do It Better" on this issue. It include an appendix of the many publications arguing against holistic scoring.)

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 don't have to *ponder* and try to make careful distinctions. After they read the paper, they just 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:

- 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'll never forget walking into my office at MIT in the 70s and finding a paper on the floor that a fearless student had slipped under the door. I had given it a B minus and carefully shown in my comments why this was exactly the right grade. But scrawled boldly across the top was this simple message. "This is a *B paper.* Fuck you.").
- 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.

(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 use four levels: poor, fair, good, excellent. But the distinction between fair and good is exactly what we shouldn't trust.)

Furthermore, 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'm avoiding the obvious bias that comes with single score holistic grades–grades that are prey to knee jerk global reactions. By using concrete objective criteria, I'm being 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 acknowledge that are saying, *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's particularly helpful for students if we give out the criteria for any given assignment *before* students have to write. It usually resuts in better essays. And we can teach better if we are willing to engage in the self-analysis of figuring what criteria we care about most and--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's 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.

A Note about Rubrics, Holistic Scoring, and Criterion Based Evalution

I see the definitive argument against holistic scoring in David McClelland's 1973 essay in the *American Psychologist*: "Testing for Competence Rather than for Intelligence." It was an argument about the problems with conventional tradition of *norm based* assessment and the advantages of *criterion based* assessment (or outcomes- or competence- 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. These descriptions seldom resemble the actual essays; they offer a kind of Platonic picture of the 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*.

The insistence on a binary yes/no result for each outcome is a problem that bedeviled the 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. 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 *in 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's hard to break them down so far that evaluation results can take the form only of a series of *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 *criterion based* evaluation that insists on articulating what is to be learned. Yet when rubric users insist on scores of 1-5 on each criterion, they fall back into the norm-based trap: fixating on fine numerical distinctions that won't hold up. But when we use rubrics that use three rather than five levels of accomplishment, e.g., *strong, okay, weak*, we are not just settling for a *compromise* between norm based and criterion based evaluation (not that there's anything wrong with compromise). I'd 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.

There was a big movement in the 70s for "competence-based" evaluation. I spent four years on an eight person research team investigating competence based programs in higher education. (We each had a site but we all visited all the sites and wrote field notes on them. See our published volume, *On Competence* [Grant] and my essay "Trying to Teach while Thinking about the End." Remarkably, David Riesman was one of our team).

For a long time it seemed as though the enthusiasm for comptence- 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 are trying to teach students to learn or be able to do--and articulate these learning goals publically and clearly enough for students to understand--*and* 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 ten or twelve 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've 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 of association of Writing Program Administrators (see WPA). Many of outcome bulldozers lead to crude and unhelpful results, but I can't 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's useless and harmful to evaluate unless the results involve *words for describing what is being evaluated*--instead of just numbers rank people as better than or worse than.

Other Common Teacher Evaluations

Before going on, let me stress the overall theme of my essay: despite all my talk of traps, pitfall, and failures to evaluate validly, in fact many teachers routinely evaluate wisely on student papers. Their evaluations may not be *wholly* trustworthy, but they are highly useful and they are at least *good enough*: the pitfalls have been largely avoided.

• **For a writing prize.** Teacher are sometimes asked to nominate a student for a writing prize. A nomination would seem to fall smack 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'd defend such a nomination as a good example of "good enough" evaluation. I can lay out more of my analysis by exploring how such a nomination relates to the three pitfalls.

With regard to the first pitfall, a single number 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 about their weightings of 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. 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 that I'm not saying that one teacher's "outstanding" will garner agreement from all readers; but at least more are more likely to call it notably good than will agree about some middling grade where the mix of dimensions is killing. (See Despain and Hilgers 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 vs. bias--the danger is even smaller. For in almost all prize situations, the teacher is not *awarding* the prize--only nominating a student. There's a committee that must adjudicate. In fact, the awarding of prizes for excellence in writing reflects a pretty valid and sophisticated understanding of how the evaluation of writing ought to work. The 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 audience has any illusion that they are looking at a "true score." They can see that the process is an *attempt* at fairness--with full open recognition of the impossibility of attaining it.

If it's a prize for a body of work, the third pitfall is avoided, since it is based on multiple texts. Even if it's a prize for just *one* essay or story or poem, there's very little pretense that the prize is a measure of ability. Everyone can see that it's the kind of prize for one performance that is so common in athletic contests. People know that this bearer of the gold medal 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 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 vs. danger.

- The need is great. For teaching and learning to go on in institutions that give credit, it's important and valid to be able to withhold credit and require them to learn enough before they go on to future courses.
- Danger. Here again, the single number verdict is not so damaging because it's at the extreme. When the teacher decides she should fail a student because of poor writing, the writing will 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 notoriously fail the test of fairness. (*I can't fail a 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") I'd argue that we must not accept such a weighty consequence as a failing grade if it's based on just *one* person's judgment about quality of writing. It isn't "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's more or less what Pat Belanoff and I set up in Stony Brook. If a teacher wanted to fail a portfolio and thus deny credit and require retaking the course again, another teacher--who didn't know the student--had to agree that the portfolio was of failing quality. (In fact, few failing grades are based on quality of writing. Usually they stem from dereliction of duty--and for deciding on that, there's no need for a second opinion. This essay is about the evaluation of writing.)

• **Eligibility to keep a scholarship or play on a varsity team**. "*Professor, you just <u>have</u> 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 certifying 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 *don't* 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 a multidimensional performances--performances right in the middle of the scale where disagreement among evaluators is virtually inevitable--and they have to present themselves as 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'd had one teacher but lost it if he'd 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 <u>some</u> way to decide on eligibility for keeping a scholarship or being on a team!* But there's 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 the main role in such teacher decisions anyway, e.g., 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've tried to justify a course grade of F, but what about all the other grades? One good thing about them is that they are almost always based on *multiple* and *different kinds* of writing, not just one text (though there are some upper level and graduate courses where teachers base the final grade on one big term paper or exam). But course grades fall squarely into the first two pitfalls: they are single number 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've defended nominations for a prize because that unilateral judgment Is simply a doorway into a collaborative judgment. I've 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 grad schools and employers. Yet the same student would likely get different grades from different teachers looking at the same work--at least with mid level quality work.

A course grade of, say, *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.*

FOOTNOTE.

The contract grading Danielewicz and I wrote about avoids all the untrustworthy mid range grading and asks us to grade on quality only for outstanding performances--and to wait till there are multiple texts to base it on. Yet it doesn't diminish the useful evaluative feedback we give to individual papers. The contract's focus on behavior lets the teacher spend very little time trying to *evaluate* writing. Instead they put their time and energy into giving writerly feedback and figuring out what student behaviors to require--what behaviors most reliably lead to learning to write better.

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 the form of a grid through which the teacher can communicate more clearly how well students have mastered the different dimensions of the course material. 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 look at 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 coures tend to have questions like these: *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?*)

FOOTNOTE. 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) of whether a question tests, say, memory, or a theoretic understanding of concepts, or an application of concepts to new material, or computational skill.

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 wouldn't actually require much more grading work, because again, *minimal crud*e 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 didn't 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 "weak" on memory--or a poor grade for the course even with "strong" for diligence and responsibility.

I don't know any college that uses a grid transcript, but it's exactly what report cards look like for children in the early elementary. Most teachers and policy makers seem to feel it is "childish" to have multidimensional grades--when in fact it is much more 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 don't 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."

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'd argue against computing a GPA for any single semester or year's performance.

But--and this highlights my nonpurist theme of "good enough" evaluation--I'd suggest that a GPA for the whole four year transcript (or two years at a community college) is valid to compute. Here's a case where this crazy single number represents so *many* judgments by different individual that it has a kind of useful, good enough believability. But it would be crucial to avoid any attempt at precision and decimal points. The limit of what makes sense would be simple integers of 1 to 5, and I'd vote for 1 to 4. However, it *would* be more trustworthy to compute how many bottom-line A's and F's a student got as a percentage of all 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 US than any other kind of writing exam. (This is the finding of three research reports: CCCC Committee on Assessment; Greenberg, Wiener, and Donovan; Lederman, Ryzewic, and Ribaudo. Cited in Greenberg, "Validity" 17.) Ed 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's a whole literature of alternatives to placement testing. The most elegant and easy one is Directed Self Placement (see Royer and Gilles). It's no longer a new and odd experiment; it's been used with satisfactory results in a wide range of institutions. But there is another alternative model (equally widely tested and used) is superior to conventional placement tests: a one credit course or a workshop that functions as a *supplement* to the regular first year writing course. In these alternatives, students who need more help to prosper in the regular first year course are identified, but not in a big test. Instead this is done by the regular teacher in the regular course classroom in the first week. Such students 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, Grego and Thompson, Kidda, and my "Writing Assessment in the Twenty-first Century.")

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 William Smith 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?" (citation)

• **Evaluation of Programs**. In the last decade or two--especially with No Child Left Behind--the amount of placement evaluations 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.

The most important step is to avoid the third pitfall. Plenty of programs do this by using *portfolios* for the before and after snapshots. Thus they are looking at multiple texts in multiple

genres that were produced on different days. It's 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's expensive to use portfolios for programmatic evaluation, but there's no way around it if we want good enough results. Some institutions deal with high costs by *sampling* students instead of trying to test 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 numbers 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've seen it happen. It's *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 improve the program--that is, improve teaching and learning. Most legislation that requires assessment requires thoughtful examination of what's 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 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 would know, of course, that they need to put together a portfolio where 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 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 weakness on different 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 *strong*s and *weak*s for each criterion. Thus they might be able to say, "Between September and June, with respect to "mounting an effective train of thinking," there were twenty percent fewer *weak*s and fifteen percent more *strong*s. They can see which dimensions seemed to show 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 a teacher giving a failing course grade for very poor writing. That is, a one dimensional score or verdict--while not wholly trustworthy--can be good enough to be used when it's at the extreme. That is, I'd 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--of 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 <u>not</u> pretending to a measure some universal Platonic skill in writing. We are only trying to measure student improvement at the kind or kinds of writing we care about at this institution. In short, our target is frankly biased, but it's the target that we care about.*"

• A vision for SAT essay tests, NAEP essay tests, state-wide 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, 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 (a) to represent the real value of the text, and (b) to represent the ability of the writer. Falling so deeply into all three traps or illusions, their scores are not even useful at the extremes. Surely we have better things to do with money than give tests that cost so much, give worthless results, and create so much unhelpful anxiety.

But I'll end the essay 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 couldn't 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've been trying to show here the courage of my pragmatism. Some "coherent" thinkers will say I've been far too purist--disallowing so much. Others will say I've been far too permissive--condoning evaluation and even *scoring* when the results 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 don't need to throw up our hands and reject all evaluation. Like Ed, 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's worth going ahead with evaluation? People need to make this calculation individually for their circumstances; there is no one right way to evaluate writing for everyone and in every context. In dealing with this puzzle, it's 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. And it's hard not to remember all the bad hires and bad articles we've seen.

Nevertheless, I've tried in this essay to use this calculus for many educational settings. I end up working my way to deciding that the following evaluations are worth making if the stake holders want them:

--Individual teachers giving multidimensional feedback or even scores on individual papers. But only three levels of quality are warranted in verdicts on 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.

--Computing a kind of rough GPA--not for a single year but for all four (or two) years. This would depend on all teachers giving a global bottom line verdicts on their grid grades. Two kinds of GPA could result: the number of *strongs* and *weaks* as a proportion of all courses; a numerical GPA limited to a scale of 4 with no decimal places. But I'm not arguing that it's better for teachers to give bottom line global verdicts for their course grades.

--Using programmatic evaluation to try to identify improvement or lack of improvement on various dimensions of writing ability--and even to identify globally strong and weak performances.

--others?

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* money for education. This calculus makes it all the harder to justify many of our current evaluations that are so dubious and that so often do great *harm* to the climate for teaching and learning.

WORKS CITED

Benesch, Sarah. "Ending Remediation: Linking ESL and Content in Higher Education." Wash DC: TESOL, 1988.

Broad, Bob. "'Portfolio Scoring': A Contradiction in Terms." New Directions in Portfolio Assessment. Ed. Donald Daiker, Laurel Black, Max Morenberg, and J. Sommers. Portsmouth NH: Heinemann/Boynton-Cook, 1993.

---. *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing*, Logan: Utah State University Press, 2003.

CCCC Committee on Assessment. *Post-secondary Writing Assessment: An Update on Practices and Procedures*. (Spring 1988). Report to the Executive Committee of the Conference on College Composition and Communication.

Despain, Larain and Thomas L. Hilgers. "Readers' Responses to the Rating of Non-Uniform Portfolios: Are There Limits on Portfolios' Utility?" *WPA: Writing Program Administration* Vol. 16, Nos. 1-2, F/W, 1992: 24-37.

Elbow, Peter. "Trying to Teach While Thinking About the End: Teaching in a Competence-Based Curriculum." *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education*, Gerald Grant, David Riesman &c, Jossey-Bass, 1979. (Printed also in my *Embracing Contraries*, NY: Oxford UP, 1986.)

---. "Writing Assessment: Do It Better, Do It Less." Lutz, William, Edward White and Sandra Kamusikiri, eds. *The Politics and Practices of Assessment in Writing*. NY: MLA. 1997. 120-34.

---. "Writing Assessment in the Twenty-first Century: A Utopian View." *Composition in the 21st Century: Crisis and Change*. Eds. Lynn Bloom, Don Daiker, and Ed White. Southern Illinois UP, 1996. 83-100.

Elbow, Peter and Jane Danielewicz. "A Unilateral Grading Contract to Improve Learning and Teaching." *College Composition and Communication*. 61.2 (December 2009): 244-68.

Grant, Gerald, Peter Elbow, Thomas Ewens, Zelda Gamson, Wendy Kohli, William Neumann, Virginia Olesen, and David Riesman. *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education*. Jossey-Bass, 1979.

Greenberg, Karen. "Validity and Reliability: Issues in the Direct Assessment of Writing." *WPA: Writing Program Administration* 16.1-2 (Fall/Winter 1992): 7-22.

Greenberg, Karen, Harvey Wiener, and Richard Donovan. "Preface." *Writing Assessment: Issues and Strategies*. Eds. Karen Greenberg, Harvey Wiener, and Richard Donovan. NY: Longman, 1986. xi-xvii.

Grego, Rhonda and Nancy Thompson. "The Writing Studio Program: Reconfiguring Basic Writing/Freshman Composition." *WPA: Journal of Writing Program Administrators*.

--. "Repositioning Remediation: Renegotiating Composition's Work in the Academy." *CCC* 46 (Feb 1996).

Herrnstein Smith, Barbara. *Contingencies of value: alternative perspectives for critical theory*. 1988

Kidda, Michael, Joseph Turner, and Frank E. Parker. "There *Is* an Alternative to Remedial Education." *Metropolitan Universities* 3.3 (Spring 1993): 16-25.

Lederman, Marie Jean, Susan Ryzewic, and Michael Ribaudo. *Assessment and Improvement of the Academic Skills of Entering Freshmen: A National Survey.* NY: CUNY Instructional Resource Center, 1983.

McClelland, D. C. "Testing for Competence Rather than for Intelligence." *American Psychologist* 28 (1973): 1-14.

Myers, Miles and P. David Pearson. "Performance Assessment and the Literacy Unit of the New Standards Project." *Assessing Writing* 3.1 (1996): 5-29.

Royer, Daniel J. and Roger Gilles, eds. *Directed Self-Placement: Principles and Practices*. Eds. Kresskill NJ: Hampton Press, 2003.

Roskelly, Hephzibah and Kate Ronald. *Reason to Believe: Romanticism, Pragmatism, and the Possibility of Teaching.* Albany NY: SUNY P, 1998.

Smith, William. "Assessing the Adequacy of Holistically Scoring Essays as a Writing Placement Technique." *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations*. Eds Michael Williamson and Brian Huot. Hampton Press, 1993.

Winnicott, D. W. "Transitional Objects and Transitional phenomena—A Study of the First Not-Me." *International Journal of Psycho-Analysis*, 1953

Council of Writing Program Administrators. "Outcomes Statement for First-Year Composition." Retrieved February February 2010 from www.wpacouncil.org/positions/outcomes.html

Added Bibliographic Note

Here are previous published essays of mine that have influenced my present thinking--arranged chronologically:

"More Accurate Evaluation of Student Performance." *Journal of Higher Education* 40 (Mar 1969). (Also in *Embracing Contraries*.)

"Shall We Teach or Give Credit? A Model for Higher Education." *Soundings* 54.3 (Fall 1971): 237-52.

"Trying to Teach While Thinking About the End: Teaching in a Competence-Based Curriculum." *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education*, Gerald Grant, David Riesman &c, Jossey-Bass, 1979. (Printed also in my *Embracing Contraries*, NY: Oxford UP, 1986.)

With Pat Belanoff., "Using Portfolios to Increase Collaboration and Community in a Writing Program." *WPA: Journal of Writing Program Administration* 9.3 (Spring 1986): 27-40.

(With Pat Belanoff) "Portfolios as a Substitute for Proficiency Examinations." *College Composition and Communication* 37.3 (Oct 1986): 336-49.

With Pat Belanoff. "SUNY: Portfolio-Based Evaluation Program." *New Methods in College Writing Programs: Theory into Practice*. Eds. Paul Connolly and Teresa Vilardi. NY: MLA, 1986.

"Making Better Use of Student Evaluations of Teachers." *ADE Bulletin* 101 (Spring 1992): 2-8. (Reprinted in *Profession 92*, Modern Language Association.)

With Kathleen Blake Yancey. "On Holistic Scoring and the Nature of Reading: An Inquiry Composed on Email." *Assessing Writing* 1.1 (1994): 91-107. (Reprinted in *Adult Assessment Forum*.)

"Ranking, Evaluating, Liking: Sorting Out Three Forms of Judgment." *College English* 55.2 (Jan 1994): 187-206.

"Will the Virtues of Portfolios Blind Us to their Potential Dangers?" *New Directions in Portfolio Assessment:* . Eds. Laurel Black, Don Daiker, Jeff Sommers, and Gail Stygall. Heinemann/Boynton-Cook, 1994. 40-55

"How Portfolios Show Us Problems with Holistic Scoring, but Suggest an Alternative." (Reprinted from *Assessment Update*. 6.4 (1994) into *Portfolio Assessment: Uses, Cases, Scoring, and Impact.* Ed. Trudy Banta. Jossey-Bass in 2004.

"Writing Assessment in the Twenty-first Century: A Utopian View." *Composition in the 21st Century: Crisis and Change*. Eds. Lynn Bloom, Don Daiker, and Ed White. Southern Illinois UP, 1996. 83-100.

"High Stakes and Low Stakes in Assigning and Responding to Writing" and "Grading: Calculating the Bottom Line." *Writing to Learn: Strategies for Assigning and Responding to Writing in the Disciplines*. Peter Elbow and Mary Deane Sorcinelli, editors. San Francisco: Jossey-Bass. 1997.

"Writing Assessment: Do It Better, Do It Less." Lutz, William, Edward White and Sandra Kamusikiri, eds. *The Politics and Practices of Assessment in Writing*. NY: MLA. 1997. 120-34.

"Taking Time Out from Grading and Evaluating while Working in a Conventional System" *Assessing Writing* 4.1 (spring 1997): 5-27.

"Changing Grading While Working with Grades." *Theory and Practice of Grading Writing: Problems and Possibilities*. Eds. Chris Weaver and Fran Zak. Albany, NY: SUNY Press 1998

"Getting Along without Grades--and Getting Along with Them Too." *Everyone Can Write: Essays Toward a Hopeful Theory of Writing and Teaching Writing*. NY: Oxford University Press, 2000. 399-421

"Directed Self-Placement in Relation to Assessment: Shifting the Crunch from Entrance to Exit." *Directed Self-Placement: Principles and Practices*. Eds. Daniel J. Royer and Roger Gilles. Kresskill NJ: Hampton Press, 2003. 15-30.

"A Friendly Challenge to Push the Outcomes Statement Further." *The Outcomes Book: Debate and Consensus after the WPA Outcomes Statement.* Eds. Susanmarie Harrington, Keith Rhodes, Ruth Overman Fischer, Rita Malenczyk. Logan UT: Utah State UP, 2005. 177-90.

"Do We Need a Single Standard of Value for Institutional Assessment? An Essay Response to Asao Inoue's "Community-Based Assessment Pedagogy", *Assessing Writing* (2006).

With Jane Danielewicz "A Unilateral Grading Contract to Improve Learning and Teaching." *College Composition and Communication*. 61.2 (December 2009): 244-68.