1

Reflections on an Explosion

Portfolios in the '90s and Beyond

Peter Elbow
Pat Belanoff

KATHLEEN YANCEY INVITED US TO REFLECT ON WHAT WE NOTICE AS WE LOOK at the portfolio explosion that has gained steady strength since we started our experiment in 1983 at Stony Brook.

First, we note that we are not assessment specialists. We have not mastered the technical dimensions of psychometrics. That doesn't mean we don't respect the field; we agree with Ed White that one of the greatest needs is for practitioners and theorists like us to talk to psychometricians. But we don't feel comfortable doing that so long as they continue to worship numbers as the bottom line. We think teaching is more important and more interesting than assessment. (Yes, teaching involves internal, informal assessment, but not external, formal assessment.) The reason we felt impelled to get deeply involved in assessment was that it began to impinge so powerfully upon teaching. The most important lesson we've learned is that people can do useful work in assessment without being on top of technical psychometrics.

The portfolio explosion has brought conferences, journal articles, essay collections, diverse experiments, research reports, and more. Portfolios are currently being used at all educational levels: kindergarten to graduate to returning adult programs. And they are being used in a wide variety of contexts: within individual classrooms, across grade levels, and within citywide and statewide assessment programs. The bulk of this activity has developed within the last eleven years. Perhaps the first thing to say is

that we can't "look back" on all of this: it's too much to see—to keep up with.

Nevertheless we are excited and bemused—and proud too.

Why the Portfolio Explosion?

The proliferation itself suggests, first, that what looks on the surface like a miraculous increase is not so miraculous after all. It makes us think of the down-to-earth interpretation of the biblical miracle of the loaves and fishes: a lot of members of the crowd had stuffed their pockets with a lot of bread and fish when they realized they were going to walk out into the desert. When it was time to eat, a lot of pockets were opened. We've discovered that many teachers, especially at the elementary level, had been using portfolios in their own quiet ways for years before we did. When we listen at the ubiquitous portfolio conferences, we hear teachers start off, "Well, in 1965, here's how I did it." We whisper to each other, "We never dreamed of portfolios that long ago!"

In short, our two essays in 1986 (and Chris Burnham's in the same year and Roberta Camp's a year earlier) brought a process and a principle to wider attention that had already existed in scattered ways. Apparently, we provided a wider conceptual scheme for an activity already underway in scattered sites. We managed to frame thinking about portfolios more consciously in terms of assessment—particularly external large-scale assessment. This process makes us think of the history of freewriting. Ken Macrorie made freewriting prominent and Peter managed to publicize it more, but as Macrorie pointed out in his historical essay (Macrorie 1991), it's an idea that had been kicking around in various forms for years and years. (For striking examples, see William Carlos Williams 1964 and S. I. Hayakawa 1962.) We can see the same thing with writing groups. Anne Ruggles Gere showed that what looked like innovation in the classroom twenty years ago was hardly news to many writers. What all of this makes us realize is that startling practical and ideological movements seldom spring from nowhere. Some catalyst draws together, foregrounds, and provides a useful conceptual framework for the growth of already existing or incipient ideas.

But if the idea of portfolios had been kicking around for so many years, what was it about 1986 and the years just following that somehow made it a catalytic situation? In retrospect, what was striking was the urgent and growing pressure for assessment, assessment, assessment; test everything and everyone again and again; give everything and everyone a

score; don't trust teachers. (This distrust was perversely reinforced at the college level in English because so many teachers were adjuncts, part-timers, or temporary.) School, district, and state administrators turned more and more to outside testing, psychometricians, and large testing agencies to ascertain and validate student learning in order to evaluate the effectiveness of teachers and programs. People began to believe that without an outside-derived number and a grade it was impossible to trust that any learning had taken place. It was in this era of growing distrust and suspicion that the steamroller movement for standards started gathering momentum. In writing, this was the era of more and more holistic testing and norming.

This greater than usual pressure for testing and bottom line, single-dimensional numbers was the matrix for a greater than usual hunger for an alternative way to assess student writing and learning. Teachers have always given grades—and no doubt will continue to do so. But never before had so many teachers and programs had to give so many single number scores for performances that are as hard to quantify as writing. For teachers who already knew how problematic such assessment was, the pressure for more of it drove them to seek assessment that was more compatible with their classroom practices. We see, in short, a dialectic process: too much pressure for X creates a striking growth of Y.

Thus the events at Stony Brook were a paradigm of the times. The faculty senate had decided several years earlier not to trust the grades given by first year writing teachers (especially graduate-assistant teachers), and therefore mandated a proficiency exam that overrode course grades: no one could satisfy the writing requirement without passing the exam—even if they got an A in the course itself. The exam was a typical, holistically scored affair. Because we so strongly resisted this system—because it made a mockery of strategies we advocated in the classroom—we were driven to find an alternative.

We were surprised and even pleased to discover that our own hunger for a different way to evaluate writing ability was echoed in so many colleagues in the widest variety of institutional settings: "You mean we don't have to do it this way? You mean grades on individual papers and writing exams are not built into the universe like gravity? You mean we're not stuck with holistic scoring?" This fertile soil led to the proliferation of portfolio evaluation. And we were lucky enough to have a forum from which to speak to a growing audience. Peter had managed to get a reputation by this time, and the discipline of composition and rhetoric had begun to establish itself as an important field that other disciplines were beginning to listen to.

A New Emphasis on Collaboration and Negotiation

Portfolios have always been useful and productive for individual teachers, but we added a new emphasis on collaboration and negotiation. What was central to our experiment was to move portfolios outside the individual classroom so that they would be read by someone else in addition to the classroom teacher. We wanted a situation where teachers had to work together and negotiate a judgment. Once we got this kind of collaborative talk going, we came to understand even more fully than before how inadequate traditional proficiency testing can be. Collaboration prompts teachers to have to articulate for others (and thus for themselves) the basis for their judgments. In the course of such articulation, we came to understand how subjective all evaluation is. No one in our program could close a door and just give grades without being influenced by other teachers.

We think we learned something important about the negotiation process. Negotiation and collaboration often break down when participants are working under too many rigid constraints. Stony Brook teachers do not have to use the conventional range of holistic scores from one to four or one to six; they just score portfolios satisfactory or not satisfactory. In addition, teachers are not obliged, in the end, to agree. What they must do is engage in the collaborative and negotiating process and listen to any differences between their judgment and that of their peers. For the vast majority of portfolios, readers do manage to agree. For a few they do not. The point is that collaboration and negotiation (and, most important, the ability actually to change your mind) work best when the situation isn't too rigid or coercive. (For more about the specifics of our Stony Brook system, see Belanoff and Elbow 1991; Elbow and Belanoff 1991.)

Collaboration and negotiation, once initiated, have a way of permeating a whole program. The evaluative process spills back into the classroom and leads to more collaboration and negotiation in teaching. If teachers have to negotiate about the end-of-semester portfolios produced by each other's students, they have a powerful incentive to collaborate and negotiate about what and how they will teach.

The collaborative dimension of portfolio assessment seems to want to spread further. Pat is currently engaged in a nationwide project in which portfolios from a variety of institutions are being read by those who are geographically quite separated (see chapter 24, this volume). Such a project engages her and her colleagues in negotiation at a much broader level. We do not yet know what the outcome of this project will be, but we already see

the value of moving collaboration and negotiation to other sites. But, since collaboration and negotiation have become such sunny words in our field these days, it is important not to forget how difficult they are and how often they fail. (For a vivid and helpful account of a problematic collaboration between a university and a school system over portfolio assessment, see Roemer 1991.)

The Effects of Portfolio Assessment on Holistic Scoring and Assessment Theory

We are excited that portfolios haven't turned out to be just another tool in the testing cabinet. Portfolios have kicked back at testing itself—helping people rethink some central assumptions and practices.

This process started when portfolios helped testers face up to a problem they had been ignoring (probably because the problem was so intractable till portfolios came along): any writing exam is inherently untrustworthy if it calls for only one piece of writing. That is, we cannot get a trustworthy picture of writing ability unless we look at various kinds of writing done on various occasions. Otherwise the sample is skewed by the genre, the prompt, the student's mood, health, and so on. Portfolios, by providing different samples written under different conditions, finally went some way towards solving this problem—giving us a better picture of what we are testing for. (This means better validity—though people now argue over different meanings for that technical term.)

But when portfolios brought this improvement, they also brought a new problem. You'd think that better pictures would lead to better rating of those pictures, but these better pictures seem to lead to more disagreement among scorers. (This is a reliability problem.) This disagreement isn't really surprising once you think about it. When scorers only have to score single samples written under exam conditions—all on the same topic and in the same genre—they have a much easier time agreeing with each other than when they score the mixture of pieces in a portfolio. In one portfolio, some pieces are stronger than others, some dimensions of writing are better than others (e.g., ideas, organization, syntax, mechanics), and in fact single dimensions or aspects of the writing may be strong in one piece and weak in another. Even one reader of a portfolio tends to get into fights with herself trying to settle on a single number score she can trust for this mixed bag. The disagreements escalate when we ask several readers with different values to agree.

Of course there is a traditional assessment technology that handles disagreement among scorers: readers are "trained" to agree in training sessions where the leaders use scoring rubrics and "range-finder" sample papers. But it turns out that this training doesn't work so well on portfolio readers. They are more ornery in their disagreements. When portfolio scorers see multiple pieces by one student, they tend to put more trust in their sense of that student, and so tend to fight harder for their judgment. In conventional, single-sample tests, they are more liable to feel, at least unconsciously, "Why fight for my judgment, when I have no evidence that this text is typical of the student's other writing—especially the writing she does in more natural writing situations." (For three recent and vivid studies of actual scoring sessions that illustrate this remarkable difficulty in trying to train portfolio scorers to agree, see Broad 1994; Despain and Hilgers 1992; Hamp-Lyons and Condon 1993. Vermont is being asked to rethink its statewide portfolio assessment procedures because the testers themselves got such low scores on inter-reader reliability.) In short, portfolios seem to kick back when people try to pin single numbers on them.

Thus portfolios have put the assessment process in a pickle. They finally give more trustworthy pictures of ability (making us realize how little we could trust those old conventional single-sample pictures), but in the same stroke they undermine any trust we might want to put in the scoring of these pictures. Of course people have been calling into question holistic scoring, grading, and single-dimension-ranking for a long time. But portfolios have finally made this critique stick better.

Still, sometimes we need a single number on a single dimension—a single "bottom line" verdict or holistic score. That is, in certain situations, we need to decide which students should be denied a place in our course or institution if we have limited resources—or denied credit, or made to repeat a course, or required to take a preparatory course. Sometimes we also want to exempt students from a course or pick students for an award or scholarship. We don't need most of the scores we normally get from holistic scoring, but occasionally we need some, and we can't just beg off and say, "Our readers won't agree because they finally see that ability is not monodimensional."

Portfolios turn out to suggest a way to deal with this problem. What about a full and rich portfolio where readers agree that most of the pieces are unsatisfactory? Are we not more than usually justified in giving this portfolio a score of unsatisfactory or failing or notably weak for this population? Similarly, what if most readers agree that most of the pieces are excellent? Are we not more than usually justified in giving a score of excellent or

notably strong, or some such label? In short, portfolios have led to the concept of minimal or limited holistic scoring.

At first glance, this procedure seems odd. For one thing it might seem theoretically scandalous to give holistic scores to portfolios at the margins and no scores at all to the rest. The process is liable to yield an unsettlingly large group of portfolios in a middle, more or less acceptable, default range. In our view, however, the real theoretical scandal comes from continuing to make all those fine-grained distinctions across the middle range: these are scores about which readers tend to disagree, and so they are simply the accident of compromise and of the value judgments unilaterally decreed by test administrators.

We are not trying to pretend that minimal or limited holistic scoring—picking out the best and worst portfolios—is truly or completely trustworthy. There is always an element of subjectivity in any evaluation process—in some cases a large element. We defend the process only because it involves making far fewer dubious judgments and making only those judgments that are most needed. In short, the principle here is the same as for surgery: since every operation carries a risk of genuine harm, we should perform surgery only when there is genuine need and a likely chance of success. Most holistic writing scores are neither necessary nor trustworthy.

Now just as it's cheaper to avoid surgery, it is cheaper to avoid all those unnecessary and untrustworthy holistic scores. Thus minimal holistic scoring recoups much of the extra cost of going from single sample assessment to portfolio assessment. With minimal scoring, most portfolios can be read in just a couple of minutes: they soon establish themselves as too good for unsatisfactory and too flawed for excellent. Scoring is faster and cheaper still if we don't need to identify top-rated portfolios. So if portfolios are used as an exit test—or if they are used for a placement procedure where students are not exempted—only poor portfolios need to be identified.

Most large-scale writing assessments are designed to sort students, not give feedback. But what if we do want to give students some feedback? What if we want to use assessment to increase learning? Portfolios come to the rescue again and show us how to give more sophisticated and useful feedback on an exam. Since portfolios are mixed bags, they invite us, by their nature, to notice differences: strengths and weaknesses within a portfolio—whether between different papers or between different writing skills or dimensions.

Once we get interested in differences rather than just single numbers, we realize that it's not so hard to communicate these differences in scoring so that the student at last gets a bit of substantive feedback from the assessment

process. For this feedback we don't need traditional analytic scoring—that elaborate process in which various writing dimensions or features are scored on a scale of four or six and these subscores are added up into a holistic score. No, it's much more feasible and trustworthy to use something simple and minimal: readers score a writing trait or dimension or paper only if they feel it is notably strong or weak. Thus there are only two scores, strong and weak, along with a third default middle range. The traits might be traditional ones, such as ideas, details, organization, clarity of syntax, voice, mechanics; or rhetorical features like finding a subject, or making contact with readers; scorers might even note individual papers in a portfolio as particularly strong or weak. (See Broad 1994, Figure 20-2 for a long list of features that readers can quickly check off as notably strong or weak while they read a portfolio—features that Broad derived from actual scoring sessions.)

Obviously, we are no longer saving time and money if we decide to give this kind of feedback to portfolios. But there is a compromise that we used at Stony Brook: we gave this kind of analytic feedback only to failing portfolios. This didn't take much time—since readers already had to read failing portfolios more carefully. And of course the failing students need this feedback most.

All of this, then, is a story of how portfolios have highlighted problems with assessment that have been lurking there all along. In particular, portfolio assessment has finally brought wider attention to the problems of holistic scoring that a number of us have been calling attention to for a long time.[1] Portfolios kick back not only at conventional holistic scoring but even at grading in general. That is, once portfolios force us to reflect on what should be obvious—namely that no complex performance can be accurately summed up in a single number because it almost always has stronger and weaker aspects or dimensions—we can see all the more clearly that conventional grades, whether on papers or for a whole course, also don't make sense. Trying to give a course grade is very much like trying to give a portfolio grade. In both cases one is trying to pin a single number on a mixed bag of performances. And so the obvious solution suggests itself: minimal or limited grading—using terms such as outstanding, satisfactory, and unsatisfactory—and adding differential notations that describe where the student did particularly well or badly. The debate about grading has tended to be binary and oversimple as though we had to choose between conventional grading and no grading (such as at Evergreen or Hampshire College). The example of portfolios shows us how feasible it is to use some kind of minimal holistic grading—along with some markers of strengths and weaknesses.

To summarize this section: portfolios have helped more people involved in assessment to acknowledge how untrustworthy it is to rank multidimensional performances along a monodimensional scale. When testing is only for placement or for identifying students who have reached a satisfactory, mere minimal holistic scoring will do. This saves money and means fewer dubious judgments. But because portfolios are mixed bags and thus invite evaluators to notice differences (things done well and not so well), they have come to suggest the possibility of scoring strengths and weaknesses.

Effects of Portfolio Assessment on Teaching

We got involved in portfolio experimentation in 1983 because of the threat to teaching posed by proficiency exams, but we had no idea of the teaching potential of the portfolio process itself. It's true that Peter, because of his three-year stint in a competence-based research project, did have a sense of some of the theoretical implications in assessment—particularly evidenced in the move from norm-referenced to criterion-referenced models of testing (see Elbow and Belanoff 1991; McClelland 1973). And Pat, during her years at NYU, had been involved in a portfolio project created by Lil Brannon as an alternative way of satisfying the writing requirement for those who failed NYU's proficiency exam. She had an opportunity to experience the difference between "scoring" a proficiency exam and evaluating a portfolio. But neither of us had any sense of how widely adaptable this portfolio creature was. And most of all we had no idea of how deeply it would reflect back on the teaching process.

Portfolios wormed themselves into everything we did. They seem to do that in many settings. They have a fruitful and supportive effect on the individual classroom, both on teachers and students. We continue to see how portfolios help teachers negotiate the conflict between the role of supportive, welcoming helper and the role of critical, skeptical evaluator. On the one hand, portfolios help separate the two roles. That is, portfolios help teachers stay longer and more productively in the supportive role, but then in turn, help them move more cleanly but less frequently into the critical role. Indeed, in a system where teachers collaborate with each other for portfolio assessment, the teaching and testing roles are separated even more since the teacher brings in an actual outside evaluator who occupies only the role of critic.

But on the other hand, portfolios help teachers unite or integrate these conflicting roles of teacher and evaluator. That is, portfolios permit us to avoid putting grades on individual papers, and thereby help us make

the evaluations we do during the semester formative, not summative. (Of course, grades on papers in a conventional course are supposed to be formative rather than summative, but because they are single number grades that go down in the grade book, both teacher and student tend to experience them as summative. This undermines the learning process.) And when teachers evaluate portfolios together at the end of the semester for summative verdicts, the fruits of their discussions tend to become internalized and help shape ongoing classroom strategies, conversation, and feedback. When all goes well, this consciousness also then seeps into students' conversations about theirs and their peers' writing. After all, self-evaluation is the strongest force for successful revision.

The important issue here for all of us in education is the way practice and theory interact and enrich one another. Our desire to replace Stony Brook's proficiency exam grew out of our acceptance of certain theories inadequately summed up as the "process movement" in composition and rhetoric. This movement led us to change our own teaching; the resulting changes in our classrooms led us to challenge a proficiency exam that contradicted how we taught the course—a course that was supposed to prepare students for the exam. By asking ourselves why portfolios seem to help our practice, we feel we can enrich our own (and we hope others') theoretical awareness of developments within the field. We will just mention here in a summary way the larger theoretical points that strike us as most important:

- Grades undermine improvement in writing because they restrict and pervert students' naturally developing sense of audience awareness.
- Writing is its own heuristic; it doesn't have to be graded to lead to learning.
- Portfolios lead to a decentralization of responsibility which empowers everyone involved.
- Teacher authority needs to be shared if writers are to have genuine authority.
- All evaluation exists within a context of literacy defined by a multitude of factors, not all of which are products of the classroom.
- Knowledge, whether of grades or of teaching strategies or of theoretical underpinnings, is a product of discussion among community members.
- Evaluation, judging, liking,and scoring are inextricably bound up together and need to be thoughtfully examined.

What's important is not so much whether we are right in these thumbnail theoretical points (and our list is not meant to be exhaustive), but the process through which practice and theory come together. Our practice led to theoretical reflections and conclusions which in turn enriched practices at many levels and sites. These enriched practices have led and will continue to lead to greater exploration of theories to explain the success (and failure) of whatever the new practices are. All of this supports our conviction that theory and practice when separated become stunted. All of us need to be both practitioners and theorists or philosophers of practice.

Potential Problems with Portfolio Use

We worry that portfolios have become a fad. Some people have jumped on this bandwagon in order to convince the public or their administrators that they're on the cutting edge. Others have trivialized or short-circuited the whole process of designing and implementing a portfolio system and thus robbed it of its peculiar ability to create a sense of ownership among those who do this planning. One way of doing this is to mandate from above procedures designed by administrators. The usual result of such short-circuiting is that those "ordered" to use portfolios just go through the motions and miss the enriching, empowering potentialities. (Again, there is an instructive comparison with freewriting: "Yes, I love to use freewriting in my teaching. My students get good grades on their freewriting, and I enjoy reading it.")

Portfolio assessment is sometimes felt as a cure-all. Indeed, because portfolio assessment is better than conventional assessment, teachers and administrators sometimes slide into treating it as desirable in itself, absolutely—thereby fueling the impulse for more assessment. So, ironically, whereas we think of portfolios as a way to hold back the assessment steamroller a bit, some people advocate and use portfolios in such a way as to accelerate that steamroller. Portfolios can actually be used in such a way as to make students feel as though every scrap of writing they ever do in a course might be evaluated—can make them feel the search-light of official evaluation shining into every nook and cranny of writing they do for any purpose.

Another uncomfortable realization: once a portfolio system is in place, it's sometimes difficult to change. If the participants have expended a lot of ingenuity, effort, and even risk, they have a big investment and may well be reluctant to change "their baby." Also, portfolio users do not always

acknowledge the inherent problems of any portfolio system. None of us should dismiss as non-serious the issues of cost effectiveness, time spent reading, the potential for abuse, and the need for constant attention to developing problems. We do not win skeptics to our side by treating these issues as easily resolved.

But one of the inherent potentialities of portfolio assessment is to invite change. For the portfolio brings more of the writing process and the teaching process—with all their idiosyncrasy and variability—right into the center of the assessment process. Teaching needs to be the dog that wags the tail of assessment rather than vice versa.

Despite the inherent potentiality for change, portfolio assessment can be administered and experienced as rigid, especially for those who come into a portfolio system after its initial creation. Currently at Stony Brook, we need to constantly prod graduate students to criticize the system and suggest new and better strategies; they look upon it as carved in stone because it was in place when they arrived. We know many resist or misunderstand the system. As one graduate student put it: "Portfolios are just the department's way of getting into our classrooms and dictating what we do." We're certain that this phenomenon is not limited to Stony Brook. We all need to seek ways for keeping portfolios vital, and up to now, a large part of their vitality is a product of the fact that those who use them are the same as those who designed them. We need to keep stressing that those who continue to use them have the power to redesign them.

For portfolios are simply the best system we currently have to assess writing while still trying not to disrupt or undermine the teaching and learning process. Surely something better will come along—perhaps an outgrowth of portfolio use. We all need to keep an open mind and welcome new developments. We cannot be chauvinistic about our baby. The many uses of portfolios described in this book are evidence of the power of portfolios to modify both thinking and practice.

Notes

1. In addition to the fact that holistic scoring is not trustworthy, it has these other problems. It gives nonsubstantive feedback: it's only a reading on a yea/boo applause meter. Worst of all, holistic scoring fuels the biggest enemy of thoughtful evaluation: judgment based on global or holistic feelings ("I like it"/"I don't like it"), rather than judgment that tries to describe and to discriminate between strengths and weaknesses. And it also feeds the pervasive hunger in our culture to rank complex performances with simple numbers—the pervasive assumption that evaluation isn't trustworthy, hardheaded, or honest unless it consists of single numbers along a single

dimension or a bell curve. Portfolios are helping more and more people realize that, as professionals, we need to convince people that evaluation isn't trustworthy unless it avoids the distortion of single numbers. Because portfolios get us to think in a more sophisticated way about the assessment of writing, more people are finally acknowledging that even a single short essay is a complex performance, and that giving it a single number is usually a distortion. (See Appendix A of Elbow, "Writing Assessment," for a long list of works criticizing holistic scoring.)