6, Numbers 1-2, Fall/Winter 1992

Readers' Responses to the Rating of Non-Uniform Portfolios: Are There Limits on Portfolios' Utility?

LaRene Despain and Thomas L. Hilgers

Portfolios are "in." Writing specialists who focus on pedagogy, assessment, and program administration, generally agree that many samples of a student's writing are preferable to a single sample. Students write differently in different genres, on different types of tasks, for different audiences, and under different circumstances. Collections of student writing yield better portraits of students as writers and promote useful faculty discussion of teaching practices. Thus, more and more writing programs use portfolios of student writing as the bases for placement and course completion and for faculty training (Belanoff and Dickson; Bishop; Elbow and Belanoff, "Portfolios"; Hamp-Lyons and Condon; Smit). Assessment specialists note that wherever generalizability is a goal of the evaluation of texts, a portfolio of samples is preferable to the single sample.

As a basis for assessment, portfolios offer enhanced validity. Validity has always been a problematic concept, more likely to exist in potential than in demonstrated reality. Even potential validity is limited by the reliability that can be achieved in rating any item or sample set.

The most frequently mentioned "successes" with portfolios have involved sets of compositions in response to a single set of prompts (Elbow and Belanoff, "Using") and speedy readings to yield a simple yes or no decision on course exit (Belanoff). When we move toward use of portfolios for more complex forms of assessment, we find ourselves moving into somewhat uncharted territory, most particularly in the area of establishing reliability in reading practices. As Sybil Carlson has noted, even one hundred samples will not guarantee generalizability; the samples still must be rated reliably. The reliability that has been achieved in judging single samples of student writing (Cooper, Diederich, Lloyd-Jones; White) has contributed significantly to the credibility of writing professionals. Retaining this credibility while using a more valid measure such as portfolios is an important goal.

What we report here--the results of our own efforts to describe readers' responses to the task of assigning scores to nonuniform portfolios of student writing--suggests that reaching that goal will not be easy. Our readers all reported satisfactions that others have reported, that is, the experiences of group training and of reading portfolios forced them to

rethink their own teaching and evaluation of writing. But the same readers' problems and reservations lead us to suggest that writing program administrators greet suggestions for the use of non-uniform portfolios with questioning restraint, especially where speed and the reliability of scoring are primary considerations.

The Structure of the Portfolio Reading Sessions

The study we report here, which might best be categorized as "action research," was part of a larger investigation of writing done in our university's standard and tutorially enriched introductory courses in writing (Despain et al.). At the end of each of ten training-plus-rating sessions, we reviewed rater logs in an attempt to deal with rater concerns in subsequent training sessions. One of our goals was to come to a better understanding of what training (or "standardizing") for raters of nonuniform portfolios might profitably involve, since models for such training are not generally available.

The raters were seven college instructors who had taught the standard introductory writing course several times. Each had also worked reliably in the rating of single essays that are part of our university's means of placing students into appropriate introductory writing courses (see Brown, Hilgers, and Marsella). During the ten sessions, the seven teachers rated 208 portfolios. These portfolios had been assembled by students from more than 20 sections of the university's standard and tutorially enriched courses in writing; while all section instructors were held to a common course policy statement, each instructor created his or her own syllabus and sequence of assignments. Each portfolio contained four pieces of writing: a description, narration, or analysis of a personal experience written in response to one or another assignment in the student author's section of the introductory writing course; an analytic essay based on reading and research, written as a course requirement; an in-class "impromptu" written by students in all sections of the introductory writing course in response to a single prompt; and an out-of-class revision of the "impromptu." Since assignments across sections were not uniform, the topics of the first two texts generally differed from portfolio to portfolio.

The first reading session established a pattern for training and rating that we followed in all sessions, with some modifications explained below. For training, readers were asked to read and then to rank a set of three randomly selected sample portfolios that had been duplicated for the training session. In the first session, no criteria were provided; readers were asked to base their rankings on "first impressions." After rankings were summarized on a chalkboard and reasons for rankings were discussed,

readers were instructed to assign A, B, C, D, or F to each portfolio just as they would assign grades to essays in a regular introductory writing course. (The decision to assign grades reflected our commitment to keep our procedures "real world"; among our teachers, few advocate not assigning traditional grades.) Discussions of reasons for rankings and for assigning grades were followed by an invitation to rerank and regrade the same portfolios. This process of discussing and rerating was repeated until a consensus on "impressions" (now second, third, or fourth impressions) and ratings had emerged.

Actual evaluation then began. Each portfolio was initially read and rated by two teacher-raters. When two readers assigned scores more than one letter-grade apart, the portfolio was rated by a third teacher, and the discrepant score was discarded.

During the hour-plus of training for the first session, we were somewhat surprised to find that readers had great difficulty in agreeing upon both rankings and grades for the sample set of portfolios, since the same readers had achieved relatively high levels of interrater reliability with single-sample evaluations. Given the open-endedness of the training and the novelty of the portfolio task, we were not surprised to find that the correlation between grades assigned by two readers on the first set of portfolios was .39, indicating a positive but rather low-level predictive relationship between rating 1 and rating 2.

Following our "action research" agenda and using reader logs and what we knew about techniques to improve reliability in other circumstances, we modified the training strategy for the second and subsequent sessions. Our overall progression was from sessions that emphasized "trusting your first impressions" to sessions in which the leader directed readers' attention to criteria in a progressively more defined scoring guide. One motive behind this progression was our desire to use reader responses to improve training. An even greater motivator was our desire to learn more about the dynamics and possibilities of assessing a single writer based on multiple samples in a portfolio. Our actions were guided by two questions. First, is it possible to overcome, in a controlled setting, the problems raised by a relative nonuniformity of portfolios from "real-world" multiple-classroom settings? Second, what methods of training seem best able to equip readers to cope with these problems?

Readers' Recurring Experiences with the Reading of Portfolios

Readers' logs from each training-and-rating session, plus our own notes on the sessions, reveal several patterns of reader experience.

1. Teacher-readers find assessment problematic when they do not know the contexts of individual essays' production. This finding has two aspects.

a. Generally, when teachers read to assess writing they are reading essays written in response to identical prompts, and they have the prompts in front of them as they read. Thus, the context for the essays is quite clear. However, because our portfolios were drawn from sections of a course that each had different assignments and because the readers read portfolios from many sections at every session, the variety in essays was great. Further, prompts for the assignments were not available for the readers.

The difficulty of reading papers without knowing the context in which they were written was a common theme in raters' commentaries from the end of every session. "Without some sense of the assignment, I don't know where to place the emphasis," wrote one. "Some teachers do not stress thesis, nor a developed intro and conclusion like I do," wrote another; "without a sense of what the assignments were, it is hard for me to grade the portfolios."

b. Readers also expressed frustration over their lack of knowledge of individual authors. In any assessment situation, one might wonder whether the papers one reads really present a fair picture of their authors as writers. But in the investigators' experiences with the assessment of single samples, such concern was usually minimal. Having a group of papers from an actual semester's work, on the other hand, seemed to heighten this concern. Reader comments often expressed strongly held views. "Norming [standardizing of ratings] in the training is a good idea, but one always grades on more than a 'norm.' Writing is never done in a vacuum, nor is it evaluated without the student's history." Another reader questioned the fairness of grading without knowledge of students' contexts: "This grading is hard because I often grade my own students according to additional variables--like improvement from the last paper. Here I have been trained to grade on the quality of the writing alone. But this session makes me more and more convinced that grading on writing alone is not fair to the students."

In the nonschool world, we often assess the quality of what we read without knowing about the context or the circumstances of the writer that spawned the work. Thus, readers' problems here may have derived not only from lack of assignment sheets but also from their associating the task with the school world and from their desire as teachers to be fair in grading. Some of the problems might be overcome by having an author's own description of contexts included in each portfolio, although this alone would be unlikely to promote reliability in grading. Other problems could perhaps be overcome by limiting features both in the portfolios and in the rating procedures that readers associate with the school world.[1]

2. Within each portfolio, different types of writing produced in different contexts often triggered readers' biases.

Comments such as those above suggest that reading multiple samples may put raters into a "real-classroom" frame of mind as they read, that is, they respond as they would in a teaching situation. Besides raising questions of context, such reading also prompts more reader bias than does reading a single sample. Many of our readers reported conflict between the group consensus on grades for sample portfolios arrived at during training sessions and their own predispositions, especially as they relate to genre, subject matter, and types of assignments.

Biases related to genre and subject matter undoubtedly affect all reading of student work, both in the classroom and in other assessment situations. Portfolios, with their multiple samples, provide multiple possibilities for reader conflict and may make readers more consciously aware of their preferences and biases. At least our readers noted their own biases openly in what are often exceptionally frank pieces of post-session writing. The desirability of countering predispositions was one of the major reasons for our moving toward more structured training methods.

For example, some readers showed a strong predisposition to value exposition more than narration or description. This was explicit in several comments (e.g., "It is hard for me to weight narrative and expository essays the same"), even after we had articulated a commitment to equal weighting of genres during training sessions 4 and 6.

Another recurring problem concerned some readers' persistent valuing of writing done in class over writing done out of class. When readers' logs made this clear, we tried to negotiate differences during training; however, even after the group had apparently agreed to assign equal weightings, evidence of some readers' biases remained.

Since no scoring guide for readers can anticipate every possible source of bias and since the sample portfolios used in training could never tap all of the possible "triggers" in the set of portfolios to be read, biases were

difficult to deal with in systematic ways. All seven of the readers commented at one time or another, some more than once, on the difficulty of suppressing their predispositions, although several suggested that training had provided some help. For example:

> The training sessions helped me to re-think my priorities and raised a number of issues, for example, the relationship of genre to the perceived "quality" of the writing.

> The training session was invaluable because first, I needed to pay more attention to instruction on the scoring guide, to give equal weight to narrative and expository papers; to see a "C" portfolio as fulfilling requirements but doing so unimaginatively . . . . I had to realize how to "screen out" of my mind any bias toward a "better" assignment.

Finally, we found that over time, even the reader who is consciously trying to avoid or compensate for biases is likely to find his or her best efforts compromised by what we now take up--fatigue.

3. Reading and assessing multiple portfolios is tiring, and fatigue is a major threat to reliability.

Fatigue is easy to overlook, but it was mentioned in readers' written post-session comments more than any other item--in 21 individual entries. It was mentioned by each reader at least once, and in conjunction with both late-afternoon and Saturday half-day readings. Comments relating to fatigue appeared after each session except Session 3, even though the average number of portfolios read in any one session was less than 20 (each read twice), or about six portfolios per reader. A typical comment: "I realize that I just read through 48 papers [including samples used in training], most good . . . . I am just now recovering from the glaze, the glaze, the glaze."

Fatigue is a concern in reading for any testing situation, just as it is early a concern in teachers' grading of student work; however, the ading of nonuniform portfolios may increase the likelihood of fatigue. is method multiplies the types of attending the reader has to do and plicates the process of rating. Clearly, this must be taken into account y decision on the possible use of nonuniform portfolios for evaluation, e it means that the project will require more time and, probably, more ders--in sum, more money--than other types of assessment.

Reactions to Efforts at Improving Inter-rater Agreement

As mentioned above, problems in reader bias showed up in the first training session as did problems with reliability. An analysis of this first session showed that readers disagreed most particularly on decisions that involved assigning grades of "C" (i.e., distinguishing between C and B and between C and D). Thus, training for the second session focused on establishing clearer understandings about this discrimination.

The sample portfolios for use in the second session were selected from those on which two readers had disagreed in the previous session. Further, although the discussion ranged freely and consensus on standards was still to be in the hands of "the community of readers," the leader directed readers' attention to aspects of texts and portfolios that seemed to have caused problems. The correlation between scores assigned in this session was .52, an improvement in agreement between raters but still far from the 1.00 correlation of total agreement.

Training for the third reading session moved even further in the same direction. First, sample packets were "made up" by combining, in some cases, essays from several actual portfolios. Thus, the training packets became "model" portfolios rather than "representative" portfolios. To accompany the packets, a more prescriptive, criterion-referenced scoring guide was prepared to describe the grades from A to F. With these changes, the inter-rater reliability ratio went up slightly to .58, again an improvement but still far from the .80 that is frequently used as the standard for minimally acceptable level of agreement in circumstances where a score has significant consequences.

Changes in training came in response not only to correlation statistics but also to observations and especially to written comments, both negative and positive, by the readers. From session 3 on, then, the training sessions focused largely on distinguishing among features of the scoring guide and on working with model portfolios. The scoring guide itself was modified twice, for sessions 5 and 8, mainly to make the language more explicit and in response to suggestions by the readers. Perhaps the most significant change occurred during session 7, when, in an attempt to increase the interrater correlation significantly, we trained readers to use the scoring guide on individual essays and then to average ratings in order to arrive at a rating for the portfolio. (See Appendix for final version of scoring guide.)

The discussion in each training session was free flowing but over time was more and more directed by the leader toward the criteria in the scoring guide and toward the model portfolios. Clearly, every change in

focus led readers a bit further from "first-impression" scoring and, possibly, from the value of the portfolio as a single whole.

That readers preferred greater direction is clear from their comments:

> The introduction of a scoring guide has helped me be more consistent . . . [it] keeps me more focused on specific grade standards and helps me avoid an unintentional curve.

> Papers of all three types tend to have characteristics from more than one grade category (as given on the scoring guide). For example, sentences may be "wonderful" while the essay as a whole is "boring." So often I wind up either averaging a B/D to a C, tending to split the grade on a given paper (A-B or B-C or C- D). The listing of items "in order of importance" on the scoring guide IS VERY HELPFUL in these difficult cases.

> The use of a scoring guide . . . and individual grading within the portfolios, all help focus the reader and, I think, make evaluation easier.

Scattered among these favorable comments are some mixed ones. These anticipate some of our reservations concerning the whole process:

> Following the priorities of the scoring guide really DOES help--provided that one does not become distracted by the 1001 strengths and weaknesses NOT dealt with by the scoring guide. Stick to the scoring guide, then; that's been hard to do, but keeps the grading from being TOO impossible.

> First let me say that the scoring guide really helps make the process of evaluating clearer. They serve as general guidelines--something to consider. I can't, however, shake the feeling that some of my decisions are arbitrary--made in what feels like a vacuum.

> The suggestion at the beginning of the third hour that we stop and re-read the sample packets C,B,D, (in that order) was extremely beneficial. After doing so,

> I could see better the fine distinctions between B-C and C-D. Due to my fatigue, I was beginning to doubt my judgment; thankfully, often checking the sample essays helped me very much in making decisions. (I'd already done some referring back to sample essays during the first two hours, when it was also helpful.) I conclude that perhaps consulting a scoring guide is not sufficient. Sample essays, particularly in the third hour, seem to be an invaluable tool in evaluating portfolios.

While it is not precisely true to say that through more focused training techniques we steadily improved reliability, it can be said that once past the rather disappointing ratio of the initial reading, we climbed to the .50 level, then to the .60 level and stayed there. This gives some indication that the direction of the training--toward more structure--worked to increase the likelihood that our readings would be consistent. The improved reliability was not, however, without a price. The standardizing scoring guide achieved a better "fit" with characteristics of individual essays than with the more varying characteristics of nonuniform portfolios. Thus, the scoring guide's use shifted focus away from the variety of expressions that is the basis for the portfolio's claim to enhanced validity.

Readers' responses to our attempt to improve reliability by moving in the direction of more structure formed an interesting pattern. Early comments on the training showed raters thinking of the complexity of real-life reading and thus resisting the discipline necessary for reading in a controlled situation.

In later sessions, this resistance seemed to disappear. Readers believed the discipline of training to be necessary, even valuable; nonetheless, some of the original resistance to the constraints resurfaced over time, at least with a couple of readers. This pattern may say something both positive and negative about the process. The comments show that readers gradually accepted the particular circumstances presented by the reading even to the point of helping shape the circumstances, but certain readers also showed a healthy resistance to a reading situation that forced a reading style on them. The comments point up the paradox in our experience with portfolio reading: Problems inherent in attempting to achieve reliability in the new and complex situation of reading several nonuniform writing samples in portfolios force a structured approach to reading that differs from the flexible reading one ordinarily finds in a real-world community of writing teachers. This is further evidence that the original intent of our decision to use portfolios--to provide a "real-world" assessment by using

a comprehensive sample of a student's writing--may to a certain extent have been compromised. Our readers put it well:

> There IS a point at which having a well-spelled out scoring guide works against you--the reading stops being holistic and becomes a labored effort to fit writing into discrete categories.

> Our going over the scoring guide so closely makes me realize something about my way of reading holistically: that it seems always to be some criterion OTHER THAN WHAT IS ON THE SCORING GUIDE that makes me feel like I am making an accurate assessment. What seemed like a great idea--to make the scoring guide more definitive--turns out to be more problematic because it is so limiting in our reading of the portfolios holistically.

Final Considerations

As with all experiences related to holistic rating of writing, ours were "unique" in that particular persons participated at particular times and in particular places in reading particular portfolios. In addition, our raters had all been previously trained to be reliable holistic raters of individual writing samples. Nonetheless, our training methods were rather conventional, and our student authors were in many respects typical of college freshmen; however, we think that some of what we learned may be helpful to writing program administrators who must make decisions on ways of assessing written texts.

Our experience taught us three things about using portfolios in situations where speed and high degrees of interrater agreement are deemed necessary. First, using portfolios takes a tremendous amount of time and energy, requires long training sessions with careful planning, and probably works best when not too many portfolios are rated by a single reader at any one session.

Second, the reading itself tends to be more unruly than the reading of single pieces of writing produced under controlled circumstances in response to identical prompts. The complexity of multiple samples keeps readers from holding in check their predispositions, even biases, regarding genres, content, and context. In addition, the lack of a clear context, both for the material to be read and for the writers of the material, presents as

yet unresolved problems for readers who see a body of work that they know was produced as part of an academic course.

Third, a paradox seems to emerge from our data. When we first started this project, a colleague rather cynically noted that inter-rater reliability depends upon a leader's imposing his or her will on raters. In a real sense, this proved true as training sessions needed to become more and more structured if readers were to reach agreement. Thus, the manipulations needed to improve inter-rater reliability may undermine the very "real-life" validity that prompted portfolio reading in the first place (for further discussion, see Hilgers & Marsella).

Those who report successful experiences with portfolios as a means of evaluating student writing appear to operate without the constraints we faced. For example, where portfolios with uniform contents serve as a "leaving exam" for a writing course, maintaining high levels of inter-rater reliability is not a pressing issue since there are other checks on the outcome (mainly the records established by the students in the course). Reading sessions involving such portfolios are conducted in a spirit of negotiation rather than with a goal of standardization.

Can the rating of portfolios with nonuniform contents be used for research projects or for program assessments that require highly reliable findings? Perhaps not, or perhaps not yet. If not yet, then we must work to discover how we can achieve the control necessary for satisfactory levels of reliability without sacrificing the validity sought through use of portfolios. If simply not, not all is lost. It may prove possible that different purposes for assessment will dictate different trade-offs between validity and reliability. Our project did not arrive at a completely satisfactory set of trade-offs, but this experience does not mean that an acceptable compromise between the validity that maintains in "real-world" settings and the reliability we seek when we want to generalize our findings about real people cannot be found. Our experience does suggest, however, that writing program administrators should carefully consider the goals and particular circumstances of assessment before they decide whether or not nonuniform portfolios and holistic ratings are the best available vehicles for the evaluation of writing.

Note

1. Elbow and Belanoff report that they began including a statement on the nature of the assignments with their portfolios at the request of their readers. They also specified what was to be included in the portfolios so their portfolios were perhaps not as varied as ours were. This might have helped in our case, but in the absence of a uniform curriculum for each section of a course, variability will always exist. Thus, inclusion of prompts, while possible in many school situations, would not only add to reading time but would also introduce another source of error--readers' (mis)readings of prompts.

Works Cited

Belanoff, Pat. "Assessment in the 1990s." *Eighth Annual National Testing Network Conference on Language and Literacy Assessment*. New York November 1990.

Belanoff, Patricia, and Marcia Dickson. *Portfolio Grading: Process and Product*. Portsmouth, NH: Heinemann, 1991.

Bishop, Wendy. "Designing a Writing Portfolio Evaluation System." *The English Record* 40 (1990): 21-25.

Brown, James D., Thomas Hilgers, and Joy Marsella. "Essay Prompt and Topics: Minimizing the Effect of Mean Differences." *Written Communication* 8 (1991): 533-556.

Carlson, Sybil B. "Meeting Measurement Standards with Portfolio Assessment." *CCCC Annual Meeting*. San Francisco, March 1989.

Cooper, Charles R. "Holistic Evaluation of Writing." *Evaluating Writing: Describing, Measuring, Judging*. Ed. Charles R. Cooper and Lee Odell. Urbana: NCTE, 1977.

Despain, LaRene, Thomas Hilgers, and Suzanne Jacobs. "A Study of Two Models for Using Tutorial Instruction in English 197L." *Report on the Educational Improvement Fund 1987/88*. Honolulu: University of Hawaii Office of Faculty Development, 1988.

Diederich, Paul B. *Measuring Growth in English*. Urbana: NCTE, 1974.

Elbow, Peter, and Pat Belanoff. "Portfolios as a Substitute for Proficiency Examinations." *College Composition and Communication* 37 (1985): 336-339.

———. "Using Portfolios to Judge Writing Proficiency at SUNY Stony Brook." *New Methods in College Writing Programs*. Eds. Paul Connolly and Teresa Vilardi. New York: MLA, 1986.

Hilgers, Thomas, and Joy Marsella. *Making Your Writing Program Work: A Guide to Good Practices.* Newbury Park, CA: Sage, 1992.

Lloyd-Jones, Richard. "Primary Trait Scoring." *Evaluating Writing: Describing, Measuring, Judging.* Ed. Charles R. Cooper and Lee Odell. Urbana: NCTE, 1977.

Hamp-Lyons, Liz, and Bill Condon. "Readers' Responses to Portfolios." *CCCC Annual Meeting.* Boston, March 1990.

Smit, David. "Evaluating a Portfolio System." *WPA: Writing Program Administration* 14.1-2 (1990): 51-62.

White, Edward. *Teaching and Assessing Writing.* San Francisco: Jossey-Bass, 1985.

Appendix

Portfolio Scoring Guide

Use this sheet with each portfolio.

1. For each student text, check the appropriate grade score in each column.

2. Compute a grade for each paper. Grade in left column should have more weight than grade in right column.

3. Average the two paper grades.

4. If the average is between grades, use the composite grade you gave to the "war" text to break the deadlock.

DESCRIPTIVE or NARRATIVE (personal experience) text:

Interest	*Sentences*
A____Captivating throughout	*A____Wonderful*
B____Holds attention	*B____Correct & efficient*
C____On the runway, but not in the air	*C____Correct, but awkward*
D____Boring and general	*D____Awkward, noticeable errors*
F____All talk	*F____Error-filled*

EXPOSITORY text:

Structure	*Support*
A____Intelligent thesis: crystal clear structure	*A____Intelligent & imaginative use of support*
B____Interesting thesis; organization clear	*B____Support substantial & well used*
C____Has thesis, but obvious and boring	*C____Minimally necessary support*
D____Thesis & structure unclear	*D____Little support, badly used*
F____No apparent thesis	*F____Filled with generalizations & undigested quotes*

["WAR" text scoring guide omitted]