ADDRESSING THEORETICAL AND PRACTICAL ISSUES OF USING PORTFOLIO ASSESSMENT ON A LARGE SCALE IN HIGH SCHOOL SETTINGS

Willa Wolcott

The use of writing portfolios—of various types and for different purposes—is gaining in popularity at all educational levels. Portfolios comprise a valuable instructional tool for both teachers and students, especially in the area of writing. Because they are gathered over an extended period of time, allowing for student revision, collaboration, and thought, portfolios serve as an appropriate vehicle for dramatizing the writing process. More significantly, they encourage students to reflect on their own growth as writers and to participate actively in critically assessing their own work.

In addition to being recognized for their instructional value, writing portfolios have increasingly been advocated as a meaningful assessment tool. In fact, Grant Wiggins suggests that portfolios comprise the model for authentic assessment, which he defines as "the performance of exemplary tasks" (703). Yet not everyone endorses the use of portfolios for assessment. For example, the National Council of Teachers of English (NCTE) Commission on Composition warns about the "bureaucratization" that results in "central offices" both telling what contents must be included in the portfolios and assigning numerical scores to portfolios for comparative purposes; as a result, the student and the teacher are removed from the process (See NCTE Council-Grams, *Portfolio Assessment Newsletter*, Sept. 1991).

To determine the feasibility of using portfolio assessment on a large scale, I received a grant from the Florida Department of Education to conduct a pilot study of writing portfolio assessment with three English classes at a local high school.[1] One class was a twelfth-grade Advanced Placement class; one, a twelfth grade regular class that included basic writers; and one, an eleventh grade regular class that included basic writers. The study lasted for six months, and the 61 students were informed about the project beforehand. During that period I met frequently with the three teachers, all of whom were highly experienced teachers and very knowledgeable about writing assessment. The project was truly collaborative: At the meetings I talked about the issues we needed to consider, and they, in turn, indicated how well certain procedures might work with their own students and with the skills they were emphasizing in the classroom. It was important that the portfolio study not interfere with the curriculum.

A major purpose of the study was to address the theoretical issues that had surfaced in a review of portfolio literature and a review of portfolio practices elsewhere. In addition to the overriding need to design portfolio programs beforehand (French), these issues included (1) the optimal degree of standardization, (2) the need for teacher and student participation, (3) the question of authenticity, (4) the suitability of different scoring approaches in terms of reliability and validity, and (5) logistical problems of time, storage, and identification of scoring contexts.

Standardization of Portfolio Contents

The issue of standardization of portfolios for large-scale assessment is a controversial one. On the one hand, many educators, including Paulson and Paulson, believe that standardization restricts the individualization of portfolios that is their real strength. On the other hand, French and Meyer, Schuman, & Angello believe that some standardization is necessary if portfolio data are to be aggregated; otherwise, they note, there is no basis for comparability. In order to facilitate the evaluations of our portfolios, we imposed some standardization both on the number of entries to be included in the portfolios and on the types of entries to be submitted as well.

First, the teachers administered in early September to each of the three classes a common, in-class topic entitled "A person who has influenced your life." The purpose of the in-class topic was not only to provide some standardization of writing assignments across participating classes, but also to ascertain what students were capable of doing in an impromptu, timed writing situation. Six months later all three classes wrote again on a common, in-class topic suggested by one of the teachers and entitled "A time in your life when you felt special." Interestingly, even though some Advanced Placement students initially objected to having to write on such generic, "bland" topics, their teacher reported that many subsequently became interested in the topics and in doing a good job about writing about them.

In addition to the in-class selections, students were asked to submit three other selections for their portfolios, including a best selection. They were also asked to provide a reflective letter in which they explained to the portfolio reader the reasons for their choice of the best selection. Finally, they were asked to submit a cover letter or form which gave background information about each paper, as well as the drafts that one paper had gone through. Thus, the ideal portfolio contained six selections in addition to a form providing the background of each entry and rough drafts of one piece.

The types of selections to be included were broadly categorized to allow for the diversity of the differing curricula, to encourage the individuality of the students, and to foster the different types of writing taking place in the classroom. Thus, the first entry was supposed to be a "narrative or descriptive or informal essay," whereas the second entry was supposed to be a "persuasive or expository or academic essay." The third entry was designated the "best piece" and could take whatever form the student wanted it to. Having a comparable number of entries that addressed a comparable range of writing types would, we hoped, eliminate some of the problems encountered by raters in other portfolio programs, such as Vermont's, in which one scorer wrote that the meaning of a rating could differ substantially if it was based on a few, as opposed to many, selections.

As it turned out, discrepancies still arose. A number of portfolios were skimpy—either because students had not been present to write on one or both of the in-class writings, or because others chose as their "best" work a piece they had already selected to fit either the informal or formal category.

The portfolios of the Advanced Placement students, in particular, were rich and deep textured, containing thoughtful academic papers on such topics as Shakespeare, the poetry of Donne, or "Hell in the Writings of Milton and Dante." Often for their "best" piece, the AP students chose a creative story, a poem or a play they had written. Even though the portfolios of the regular students typically did not contain as many academic essays as the AP students' portfolios, strengths appeared in the portfolios of some regular students as well. Indeed, their academic essays on such topics as "Heroism in Beowolf," "Macbeth," or "The Need for AIDS Testing," when juxtaposed against the informal writings of the students, allowed readers to see where the individual students' strengths or weaknesses lay.

As can be seen then, designating the number of entries and the broad types of writing to be included proved helpful in providing some comparable basis for evaluating the portfolios—even though unevenness continued to occur in the portfolios.

Scoring Procedures

Other central issues of portfolio assessment deal with scoring methods. One decision involves whether the entries will be scored individually or whether the portfolio will be evaluated in its entirety; in a number of programs, for example, such as those of Miami University of Ohio and the public schools of

Cincinnati, Ohio, portfolios are scored as a whole. However, other educators, such as Peter Elbow and Richard Larson, argue that the complexity of portfolios belies the giving of a single holistic or summative score. Another decision entails who will do the actual scoring—internal scorers who have the students in their own classes or external scorers who have had no real contact with the writers of the portfolios. For example, the practice followed by Vermont and Kentucky is to have teachers evaluate their own students' portfolios and then to have five of the portfolios selected at random and sent to an external committee for an independent scoring as verification. In several college programs in which high stakes are involved for the individual students, the evaluators have had no previous contact with the portfolio writers.

For the pilot study, eight writing instructors—including the participating teachers and myself—gathered one weekend to score the portfolios analytically and then holistically. I initially chose analytic scoring as the primary method because I felt that a single, holistic score might be difficult to assign in view of the variety of discourse forms contained in the portfolios, and I also wished to provide feedback to the students participating in the study. To do the analytic evaluations, the scorers used a scoring sheet (See Table 1) to rate each entry in the portfolios on nine different writing elements that dealt with rhetorical issues and with grammatical and mechanical concerns. One item measured the extent to which the reflective letter showed self-assessment skills, while an optional "bonus" category also enabled the scorers to reward exceptional creativity, voice, originality, or humor in the portfolios. Each element was rated on a scale of 1-4, with 4 being the highest. Prior to the actual scoring, the scorers trained together using two portfolios from the two regular classes. After reviewing written descriptors, the scorers each rated the portfolios independently and compared results, discussing their interpretations of the key whenever conflicts arose. The teachers participating in the pilot study generally refrained from scoring any of the portfolios written by their own students, because they felt that they lacked the necessary objectivity and tended even to become somewhat critical of their own students.

Scoring the portfolios analytically took an average of 15-20 minutes, with some portfolios, such as those from the Advanced Placement students, taking well over 30 minutes apiece. Table

2 depicts the results of the analytic scoring. Not unexpectedly, the AP class received consistently high average scores, with every portfolio receiving an average score of at least a 3 and several nearly achieving a perfect score of 4. A substantial range occurred for the students in classroom Z, with some students averaging over a 3 and some averaging below a 2. A range also occurred for students in classroom Y. This range is not surprising in view of the presence of some basic writers in regular classes. That no student's average score fell below a 1.5 could be significant in showing the value of revising as a tool that can help even the weakest writers to improve; however, it must be noted that no penalties were assigned in this scoring for what might be missing from a given portfolio. That is, students' scores were averaged on the basis of what they actually submitted, rather than on the basis of what they should have included.

As a gauge of interrater reliability, 19 (30 percent) of the portfolios were given a second analytic scoring on the following day. These portfolios were selected at random, with at least five portfolios coming from each classroom. When alphas were run on the average score that each of the two readers gave on the sample of portfolios scored twice, the coefficient alpha was .83, denoting a reasonable interrater reliability rate for analytic scorings.

Most of the differences arose from contiguous scores. In 9 of the 19 portfolios the differences were consistently higher in both the rhetorical and the grammatical areas for one of the two readers, suggesting that in each case, one reader may have had the tendency to score more leniently or more harshly than the reader against whom she or he was paired. Still another contributing factor to the contiguous scores was the use of bonus points, which only one of any given two readers assigned in 5 of the 19 portfolios. In 11 of the 19 portfolios, splits—or non-adjacent scores—occurred on a few of the 46 specific items within the portfolios. Splits occurred *across* more portfolios on the rhetorical items, especially those items addressing thesis, focus, and thoughtfulness on content. However, the total number of splits *within* any given portfolio tended to be higher in the area of mechanics and grammar.

On the second day the scorers evaluated some portfolios holistically; the scorers did not rate holistically any of the portfolios they had previously scored analytically. Using a scale of

1 to 4 with 4 being the highest, the scorers assigned a single summative score that reflected the overall quality of the portfolios. Then, after giving the single score, they rated the overall quality of such key elements within the portfolios as development, content, sentence structure, and mechanics. They could also mark bonus categories for creativity and voice; and they could, if desired, add an optional comment. An important reason for including the individual ratings of key elements of the overall portfolios was to provide some feedback to students, a feedback which is lacking in holistic scoring and which is one strength of an analytic scoring. (In fact, portfolio advocates Paulson and Paulson have recommended that some combination of holistic and analytic scoring be done.) Table 3 illustrates the holistic scoring sheet.

Despite the diversity of discourse forms reflected in the portfolios, the scorers were readily able to assign a single, holistic score to reflect the overall quality of the writing. When alphas were run to determine the rate of interrater reliability, the coefficient alpha was .826, a figure comparable to the coefficient alpha for the analytic scale. Scorers spent approximately 6 minutes per portfolio in the holistic scoring, although the thick portfolios of the Advanced Placement students required more time. Precisely because of the high quality of many of the AP portfolios, all the scorers agreed that a broader scale, such as a 6-point scale, would be necessary to reflect the range of writing they encountered. The scorers experienced little difficulty in marking their overall impression of the individual elements in the portfolios, and several wrote optional comments on the score sheets. Several scorers suggested including such features as diction, grammar, and usage in the individual ratings.

In order to see how well the original analytic scores correlated with the holistic scores of the 19 portfolios, the Pearson Product Moment Correlation was run; the Spearman Rank Order Correlation was also run to see how comparably the two scoring approaches ranked student papers. The Pearson correlation was .77, and the Spearman rank order correlation was .71. While not high, both correlations seem reasonable given the small sample size of 19 and the compressed scoring scale for the holistic scores—in which basically, on the papers selected at random from the analytic set only scores of 2 through 4 were given.

The in-class topics, both of which had drawn on personal experience, proved accessible to everyone. Although students in all three classes tended to choose similar subjects to discuss, the stronger writers in the three classes often provided a fuller context for the influential people or events they wrote about, and they also demonstrated more insight into the actual meaning the person or the event had had on their lives. One teacher participating in the study noted that she found it both interesting and helpful to see what the students in the other two classes had done with the same topics.

In all three classes, over 50 percent of the students showed improvement in their analytic scores from the first in-class topic to the last. (See Table 4). Even though these figures suggest that the majority of students in all classes showed growth, such a conclusion must be cautiously made. That is, in-class writing—with its time restrictions and its lack of resources—negates much of the emphasis most writing classes put on revising, on multiple drafting, and on collaborative learning. Moreover, as the students who wrote on only one in-class topic were eliminated from analysis, the extent of any growth to be noted is limited to a portion of the pilot group as a whole. At the same time, it is encouraging to note that all three classes improved in the rhetorical areas, as well as in the area of mechanics and grammar. This finding counteracts the criticism of Knoblauch and Brannon that an emphasis on growth tends to focus on minor and measurable skills rather than on less measurable traits. (See Sommers).

Authenticating Student Writings

The in-class writings also served to authenticate the extent to which students have actually composed their own portfolios, even though authorship was never of real concern in this study since high stakes were not involved. However, this issue is a potentially troublesome one, especially in those situations in which students do have a lot at stake. Other means for authenticating ownership in this pilot study were the reflective letters and the multiple drafts that were required for one portfolio entry.

Reflective Letters

The reflective letters in the portfolios were scored differently from the other entries; that is, they were rated on the degree to which they reflected insight on the individual students' part into their own writing capabilities and performances. Not unexpectedly, two-thirds of the students in the Advanced Placement class received the highest scores possible for the insight their self-evaluations revealed. The analyses of these students often underscored their thoughtfulness, creativity, personal insight, and occasional, whimsical humor. Student 14X chose a serious work as an example of his best work for the following reasons:

> I chose 'Elements in Shakespearean Comedy' as my best piece because I felt that it was my best analytical piece. I feel that I presented a clear thesis, developed it and proved it with parts from the play ----. This paper is much better than some that I wrote at the beginning of the year, which were unclear and unorganized. This piece reflects my improvement in those two areas and in understanding Shakespeare. There are some syntax and documentation errors but I have learned from those mistakes.
>
> (*Rating of 4*)

For some of the students in the regular classes, the reflective letters seemed to be difficult to write, and one of the teachers observed that a few of her students did not seem at all interested in the "why" of their choices. Nevertheless, in both regular classes, over half the students received at least upper-half scores for their self-assessment skills. For many of the students in these classes, their "best" piece was their "favorite" piece. Thus, a number of students talked honestly about the problems in certain papers which they nevertheless rated as their best for personal reasons or because they liked the topic. The reflective letter of student 7Y is typical of many such students' self-assessments.

> After looking through all of my writings, I chose the one that I thought was my best. It was a hard decision but the one I chose is the best example of my writing ability. The piece that I chose is about Benjamin Franklin's virtues, and how the world would be if everyone followed through on his virtues and used them as guidelines for their lives. I think this is my best writing because I demonstrate my ability to linger and link paragraphs and the way I express my beliefs and ideas and opinions. There's always room

> for improvement in my writings, but I thought that this one was the best example of me as a writer.
>
> (*Rating of 3*)

Despite the difficulties students seemingly experienced in writing the reflective letters, the self-evaluation such letters necessitated remained an important part of portfolio assessment. This study showed, as has other research (see Camp and Levine; Howard), that such awareness is not readily developed; hence, students need to be given several opportunities to reflect about their writing and to determine why some selections are better than others.

Logistics

The study also suggested that the logistics of portfolio collection and scoring should be standardized and simplified as much as possible. For example, requiring students to use a common cover form and to label the kind of entry that each submission represents would eliminate a potential source of confusion for the readers. In addition to making the portfolio entries easier to score, cover letters that explain the context of each submission also provide a fuller picture of each selection and thereby serve to authenticate the authorship of the portfolios. As Gentile points out, teacher notes that explain the background of the assignments are also helpful. Furthermore, to protect the privacy of students and teachers alike, the students' names on the portfolios should be masked and coded prior to a scoring, and teachers' grades or summative comments should also be covered.

Conclusion

The portfolios that appeared in the pilot study provided a good, indepth picture of students' writing—of their strengths and weaknesses, their struggles and potential, and the progress they had made during the term. The portfolios revealed, moreover, the students as individual people.

The variety of abilities, discourse forms, and topics that were reflected in the portfolios did not present an insurmountable challenge for the scorers. Rather, the results suggested that scorers were able to assess the quality of the writings quite reliably given the complexity of the task. But what must be stressed is that the scorers, all of whom were highly experienced at scoring, underwent training for these specific scoring tasks.

Writing portfolio assessment on a large scale thus seems to be feasible provided that teachers are involved throughout the entire process and that safeguards are used to allow for individual creativity, for varying curricula, and for common requirements. Admittedly, the pilot study was small and included only three classes of varying abilities in one school, all of which had participated in the Florida Writing Enhancement Program. However, both the potential strengths and the potential weaknesses of portfolio assessment that surfaced in this study seem relevant to a number of other school systems.

Two key factors necessary for making portfolio assessment feasible are balance and participant involvement. That is, outside writings need to be balanced with some in-class work, just as informal writing assignments need to be balanced with academic essays. Standardized requirements need to be balanced with opportunities for individual writings, and student occasions for self-selection need to be balanced with guidance from the teacher. Finally, even though holistic scoring seems to be the most effective means of evaluating portfolios, that approach needs to be balanced with some analytic feedback to students or to schools. Again as the literature review notes (see Rigney) and as practice has shown, training in scoring portfolios is essential.

In addition to providing balance, a portfolio assessment procedure needs to ensure the involvement of students and teachers alike throughout the process. Portfolios imposed from the outside run the risk of being perceived as a time-consuming burden on everyone in the classroom (Cooper). However, portfolios that involve the teachers, as well as the students, in all phases of the portfolio program—from determining the nature and number of entries, to devising guides for reflective questions or generating descriptions of score levels—can become, as this pilot study suggested, a learning experience for everyone. When, as Camp suggests, portfolios are seen as an opportunity to demonstrate not only what has been learned but also what remains to be learned, then portfolios will have strengthened what should be an integral link between assessment and instruction.

WORKS CITED

Camp, R. & Levin, D.S. "Portfolios Evolving: Background and Variations in Sixth-through Twelfth-grade Classrooms." In P. Belanoff and M. Dickson (Eds.), *Portfolios: Process and Product*. Portsmouth, NH: Boynton/Cook, 1991. 194-205.

Camp, R. "Thinking Together about Portfolios." *The Quarterly of the National Writing Project and the Center for the Study of Writing, 12*(2). (1990): 8-14, 27.

Cooper, W. (Fall 1990). "Editorial." *Portfolio News* (Fall, 1990): 1.

DeWitt, K. (April 24, 1991). "Vermont Gauges Learning by What's in Portfolios." *The New York Times Education*. (April 24, 1991): A23.

Elbow, P. & Belanoff, P. "State University of New York at Stony Brook Portfolio-based Evaluation Program." In P. Belanoff and M. Dickson (Eds), *Portfolios: Process and Product*. Portsmouth, NH: Boynton/Cook, 1991. 3-16.

French, R. *Issues and uses of student portfolios in program assessment*. A paper presented as part of a "Symposium examining assessment strategies of the Next Century Schools Project" at the AERA, Chicago IL, 1991.

Gentile, C. *Exploring new methods for collecting students' school-based writing: NAEP's 1990 portfolio study*. Princeton, N.J.: Educational Testing Service, 1992.

Howard, K. "Making the Writing Portfolio Real." *The Quarterly of the National Writing Project and the Center for the Study of Writing, 12* (1990): 4-7, 27.

Koratz, D. "New Report on Vermont Portfolio Project Documents Challenges." *National Council on Measurement in Education Quarterly Newsletter, 1* (January 1993): 4.

Larson, R. L. "Using Portfolios in the Assessment of Writing in the Academic Disciplines." In P. Belanoff and M. Dickson (Eds.), *Portfolios: Process and Product*. Portsmouth, N.H.: Boynton/Cook, 1991. 137-150.

Meyer, C., Schuman, S. & Angello, N. "NWEA White Paper on *Aggregating Portfolio Data*." Lake Oswego, Or.: Northwest Evaluation Association, 1990.

NCTE Council-Grams. "Portfolio Assessment: Will Misuse Kill a Good Idea?" *Portfolio Assessment Newsletter, 3* (September 1991): 1.

Paulson, F. & Paulson, P. "The Ins and Outs of Using Portfolios to Assess Performance." An expanded version of a paper presented at the National Council on Measurement in Education in Chicago, Il., May, 1991.

Rigney, S. "Vermont Responds." *National Council on Measurement in Education Quarterly Newsletter, 1* (January 1993): 4.

Sommers, J. "Bringing Practice in Line with Theory: Using Portfolio Grading in the Composition Classroom." In P. Belanoff and M. Dickson (Eds.), *Portfolios: Process and Product*. Portsmouth, NH: Boynton/Cook, 1991.

"This is my best." The report of Vermont's Writing Assessment Program (pilot year 1990-1991). Montpelier, Vt. Vermont Department of Education.

Wiggins, G. (May 1989). "A True Test: Toward More Authentic and Equitable Assessment." *Phi Delta Kappan* (May 1989): 703-713.

Table 1.
Scoring Sheet for Portfolios

Student I.D. ______________________________

RATER NUMBER ______

	Narrative or Descriptive or Informal Essay	Persuasive or Expository or Academic Essay	Best Piece	First In-Class Writing	Second In-Class Writing	Reflective Letter	Total Score
1. The paper reflects either a clear purpose or a clear thesis that is stated or implied.						▨	
2. The paper seems focused and organized.						▨	
3. The paper is fully developed.						▨	
4. The paper reflects a thoughtfulness of content.						▨	
5. The word choice is appropriate for the subject at the secondary level.						▨	
6. The sentence style is clear and varied, with appropriate sophistication for the secondary level.						▨	
7. The paper reflects control of sentence structures. (Fragments, run-ons, and tangled syntax are avoided.)						▨	
8. The paper reflects control of usage. (Errors in subject/verb agreement, pronouns, and dialect are avoided.)						▨	
9. The paper reflects overall control of punctuation and spelling.						▨	
10. The reflective letter shows self-assessment skills	▨	▨	▨	▨	▨		
BONUS: The _______________ (creativity, originality, humor, voice) in this portfolio is noteworthy	▨	▨	▨	▨	▨	▨	
OVERALL WRITING SCORE							

KEY

Very much so	=	4
To some degree	=	3
Not very much	=	2
Not at all	=	1
Not Available	=	NA

	Yes	No	Some-what
Knowledge of the writing process is reflected in the drafts of one paper			

Sum of all papers _______

Highest score possible	=	188
Lowest score possible	=	46

Table 2.
Students' Average Analytic Scores

	Class X	Class Y	Class Z
Scores from 3.5 - 4.0	13 students (72%)		1 student (4%)
Scores from 3.0 - 3.4	5 students (28%)	2 students (11%)	4 students (17%)
Scores from 2.5 - 2.9		8 students (44%)	14 students (58%)
Scores from 2.0 - 2.4		7 students (39%)	3 students (13%)
Scores from 1.5 - 1.9		1 student (5%)	2 students (8%)
Scores from 1.0 - 1.4			

Table 3.
Prototype Holistic Scoring Sheet

RATER NUMBER: _____ Student I.D. _________

OVERALL PORTFOLIO SCORE □

OVERALL RATINGS OF FEATURES

	Excellent	Very Good	Good	Fair	Below Average	Poor
Thoughtfulness of Content						
Development/Organization						
Diction						
Sentence Structure/Style						
Grammar and Usage						
Mechanics						
Self-Evaluation Skills						
OPTIONAL Creativity						
Voice						

OPTIONAL:
General Comments: ______________________________

____

____

Table 4.
Students' Performance on the In-Class, Pre-Post Topics

Class	Number of Students who took Pre- and Post Topics	Number of Students who Improved	Average Number of Points Improved	Number of Students who declined	Average Number of points declined	No Change
X	14	8 (57%)	4.3 Points	3 (21%)	3.7 points	3*
Z	17	9 (53%)	3.6 points	6 (33%)	2.2 points	2 (22%)
Y	13	7 (54%)	5.6 points	5 (38%)	2.6 points	1 (8%)

*Two students with perfect analytic scores on the pre-essay encountered a ceiling effect