The Challenges of Rating Portfolios: What WPAs Can Expect

Jeffrey Sommers, Laurel Black, Donald A. Daiker, and Gail Stygall

Convincing proof of the profession's growing interest in portfolio assessment was offered by the 1992-93 CCCC proposal forms; for the first time, "Portfolio Evaluation" was listed as one of the major subject areas for proposed presentations. Portfolio assessment deserves such endorsement. Proponents have argued that it is a more valid method of assessing student writing (Camp; Elbow, "Virtue"; Hamp-Lyons and Condon; Huot, "Literature"); that it provides opportunities for faculty development (Belanoff and Elbow); that it upholds program standards (Ford and Larkin); and that it provides local control over writing assessment (Black et al.; Condon and Hamp-Lyons; Huot, "Beyond"; Tirrell). Yet despite a growing consensus in support of portfolio assessment over former single-sample assessment programs, much work remains to be done, particularly in scoring portfolios.

This article grows out of work we have done with the Miami University Portfolio Program, which uses portfolios for placing new students. Students who wish to participate—the program is entirely voluntary—submit four specified pieces of writing in June, prior to their first college semester. The required prose pieces, which total no more than twelve typewritten pages, include the following:

- a reflective letter addressed to the raters
- a narrative or descriptive piece
- an explanatory or exploratory or persuasive piece
- a textual analysis

Students may earn advanced placement and credits based upon their portfolios. Excellent portfolios earn students six credits and placement out of Miami's composition sequence; and very good portfolios earn students three credits and placement into a one-semester advanced composition course. Remaining students are placed into the regular two-course composition sequence.

Portfolios are holistically scored by Miami's faculty, and what we have learned is that the problems portfolios present in holistic scoring

situations merit careful thought for writing program administrators who oversee portfolio programs. Some difficulties of holistic assessment (Charney) are exacerbated by the portfolio approach, while some difficulties are unique to portfolio evaluation. While we propose to discuss some of these difficulties in this article, we cannot always offer firm solutions to the problems we raise. The purpose of this article is to stimulate on-going discussion about holistic scoring of portfolios in hopes of resolving some of its attendant difficulties. We will first look at issues of reliability in scoring and then focus on problems in redefining holistic rating in portfolio rating situations.

Problems in Reliability

The most obvious difficulty in rating portfolios is doing so reliably. Over the past two decades, we have learned much about holistic scoring techniques. We can be confident that we know how to score single samples of student writing reliably, that is, with scores from multiple raters that correlate with one another at an acceptable level. Part of the methodology of holistic scoring is devising a scoring guide that is, in some sense, an attempt to describe a range of ideal responses to the prompt. Essays that come closest to that description earn the highest scores; however, portfolios present a crucial difference that complicates the rating situation. Instead of a single, pretested assigned topic, portfolios allow students to make choices of several different pieces. Portfolio scoring guides simply cannot describe ideal portfolios with the same precision as a scoring guide for a single-sitting assigned topic essay. In fact, portfolio scoring is "messy" (Belanoff and Dickson; Yancey). Even though our portfolio requires students to write in specific genres, the contents of the portfolios still differ significantly. Of necessity, the scoring guide has had to be broader. The textual analysis alone offered a myriad of possibilities because students were advised that they might analyze a published text, the text of classmate, or even a text they had written themselves. Thus, one portfolio might include an analysis of Shakespeare's *Hamlet*, another might analyze Faulkner's *The Hamlet*, and a third might analyze the student's own poem entitled "Working at the Hamburger Hamlet." Similar variation has also occurred in the other three pieces of writing.

Thus, portfolios clearly give students more power in the assessment situation because they get to choose their topics, which leads to a number of claims that portfolio assessment is a valid assessment, one that gives us a good look at students' writing ability (Elbow, in Belanoff and Dickson, "Foreword" ix-xvi; see citations on the first page of this article). But giving students more power reduces the degree of control that administrators can exercise in single-sitting timed essays assessment. Thus, it becomes reasonable, and important, to ask whether portfolios can be scored reliably in large-scale scoring sessions. Our experience suggests that the answer is yes.

Rating Portfolios Vs. Rating Single-sitting Essays

Prior to the Miami University Portfolio Program, the Department of English administered a single-sitting proficiency essay for placement of incoming students. In the first transitional year of our portfolio program, we offered students two methods of achieving advanced placement: by submitting a portfolio or by writing an essay in two hours on campus in response to an assigned prompt. Both portfolios and timed essays were scored on a 1-6 scale. Our data suggest that rating portfolios can be as reliable as rating essays. [1]

Table 1 demonstrates that our portfolio scoring[2] was as reliable as our timed essay scoring. The numbers are remarkably close; 85.5% of first and second readers of the essays recorded scores no more than one point different from one another, while 85.8% of first and second readers of portfolios recorded scores no more than one point different from one another. In other words, if an essay or portfolio received a score of 3 from a first reader, 85% of the time the second reader recorded a score of 2, 3, or 4.

Table 1. Differences Between First and Second Readers (1-6 scale)

	Timed Essays	Portfolios
0	39.4%	45.1
1	46.1	40.7
2	12.4	12.3
3	1.9	2.0
4	.1	0

In situations where the first two readers issued scores differing by 2 or more points, we had a third reader assign a score (See Tables 2 and 3).

Table 2. Differences Between Second and Third Readers (1-6 scale)

	Timed Essays	Portfolios
0	34.4%	36.4
1	42.0	39.0
2	17.2	22.0
3	5.1	2.5
4	1.3	0

Table 3. Differences Between First and Third Readers (1-6 scale)

	Timed Essays	Portfolios
0	24.8%	39.8
1	47.1	44.9
2	24.8	7.6
3	2.5	6.8
4	.6	.8

For those essays and portfolios that required a third reading, 76.4% of second and third essay raters differed by no more than one point in their scores, while 75.4% of second and third portfolio raters differed by no more than one point in their scores. Where the percentages shift slightly, the shift favors portfolio rating. Comparing first and third readers, 84.7% of portfolio raters differed by less than one point in their scores, but only 71.9% of essay raters were that close in their scores. While it may be too optimistic to suggest that reliability may increase by using portfolios, it seems fair to assert that reliability in portfolio rating can compare favorably to reliability in single-sitting essay rating.

Rater Survey Data

We also administered a survey to our portfolio raters. These follow-up surveys of raters suggest that the reflective letter has a potentially powerful influence on rating. In fact, we wish to argue that these reflective letters help us achieve reliability in rating. Here are the instructions provided to students for the reflective letter:

> This letter, addressed to Miami University writing teachers, introduces you and your portfolio. It may describe the process used in creating any one portfolio piece, discuss important choices in creating the portfolio, explain the place of writing in your life, chronicle your development as a writer, assess the strengths and weaknesses of your writing, or combine these approaches. Your letter should provide readers with a clearer understanding of who you are as a writer and as a person.

One rater noted, "I found the reflective letter to often be the most interesting part of the packet, not only because of what it revealed of the individual but because of what it showed about the writer's attitude towards their own work. What a fascinating range of boastfulness, self-effacement, wit, and rambling." Another commented, "The reflective letter fascinates me. It appears to be the place where the student establishes his/her authority as writer; positions the reader and writer." A third rater echoed the second: "I liked those reflective letters and narratives which situated the writer and his or her writings best." The tenor of these comments, and many others like them by other raters, suggests that raters feel better prepared to read the remainder of a portfolio after reading the reflective letter. Our research into the reflective letters continues; at this point, however, we are already convinced that they affect the rating situation in a powerful way. Our speculation is that they affect reliability in a positive manner because they bring the personal back into the scoring situation.[3]

Another factor that might affect reliability is the longer reading time required for portfolio assessment. Proficiency exams usually require 2-3 minutes to read, while the Miami portfolios require 7-10 minutes each. Yet our initial research suggests that fatigue is more likely to affect proficiency exam raters than portfolio raters. In both rating sessions, raters were given 10-minute breaks every hour, similar lunch breaks, and a rating day that began and ended at similar times. Yet it was proficiency exam raters who chose to comment on fatigue: "The simple fatigue that results from hours of reading similar handwritten essays takes a toll on fairness, I'm afraid. I believe I was more fair to the portfolios, with their variety and comparative ease of reading." "Fatigue—by late afternoon—affects judgment. Only *really* good 5's or 6's jump out and get proper scoring by now." "I think that I have done my best, but I wonder if getting tired, etc. as the grading goes on distorts my fairness and accuracy." "An obvious advantage in my eyes at the end of the essay session is the possibility of variety in subject matter. This [rating proficiency exams] is getting pretty tiresome."

On the other hand, portfolio raters did not volunteer comments about fatigue. Instead, they commented on the stimulation of reading portfolios that differed from one another, unlike proficiency exams that all address

the same topic. The "messiness" of portfolios is a product of their variety, which to some raters turns out to be an advantage to rating portfolios. Raters also commented on the ease of reading typewritten work instead of crabbed handwriting. A sampling of comments from surveys follows: "Reading the portfolios is obviously more work, but I think the portfolios are a more accurate assessment of a student's writing ability." "As the day wore on, I was more interested in reading portfolios than the timed essays." "The variety of topics in portfolios makes it much easier to stay fresh and perhaps even fairer."

The surveys also asked raters to agree or disagree with a series of statements about the rating (see Table 4).

Table 4. Response to Survey Statements by Portfolio Raters (n = 34):

Statement A: "I believe that a writing portfolio, such as those I scored today, fairly and accurately reflects a student's writing ability."

Strongly Agree:	17(50%)
Agree:	15(44%)
Not Sure:	1(3%)
Disagree:	1(3%)
Strongly Disagree:	0(0%)

Statement B: "I believe that the scores I assigned to portfolios today were fair and accurate."

Strongly Agree:	9(26%)
Agree:	22(65%)
Not Sure:	3(9%)
Disagree:	0(0%)
Strongly Disagree:	0(0%)

The portfolio raters' responses show a clear pattern; the raters had confidence in the portfolio as a measurement of student writing ability and in themselves as accurate portfolio raters. While the raters' belief in the principle of portfolio rating was stronger than their own self-assessment after completing their rating, nonetheless not a single rater felt her or his rating had been inaccurate.

Proficiency essay raters were asked to respond to similar statements in a second survey (see Table 5).

Table 5. Response to Survey Statements by Proficiency Essay Raters (n = 12):

Statement A: "I believe that a proficiency examination, such as those I scored today, fairly and accurately reflects a student's writing ability."

Strongly Agree:	0(0%)
Agree:	1(8%)
Not Sure:	6(50%)
Disagree:	4(33%)
Strongly Disagree:	1(8%)

Statement B: "I believe that the scores I assigned to proficiency examinations today were fair and accurate."

Strongly Agree:	1(8%)
Agree:	8(67%)
Not Sure:	1(8%)
Disagree:	2(16%)
Strongly Disagree:	0(0%)

The responses from this second group of raters were noticeably different. As a group, they are not nearly as confident in the single-sitting essay as a measure of student writing ability as the portfolio raters were about portfolio assessment. It should also be noted that even though essay raters and portfolio raters responded similarly to the statement about their own rating in the categories "strongly agree," "agree," and "not sure," some essay raters doubted their own fairness and accuracy in a way that no portfolio raters did. The sample, however, is fairly small, and we want to resist making too much of a point here. We can confidently state, though, that the perceptions of the portfolio raters themselves support the notion that portfolio rating can be as reliable as essay rating.

Further support for that contention can be found in a third survey we conducted. Sixteen raters scored both portfolios and essays, and we administered a comparative survey to them (see Table 6).

Table 6. Response to Survey Statements by Raters of Both Portfolios and Proficiency Essays (n = 16):

Statement A: "I believe that portfolio assessment is most appropriate for awarding credit and advanced placement in college composition."

Strongly Agree:	13(81%)
Agree:	3(19%)
Not Sure:	0(0%)
Disagree:	0(0%)
Strongly Disagree:	0(0%)

Statement B: "I believe that the scores I assigned to portfolios last Thursday were more fair and accurate than those I assigned to proficiency examinations today."

Strongly Agree:	7(44%)
Agree:	5(31%)
Not Sure:	3(19%)
Disagree:	0(0%)
Strongly Disagree:	1(6%)

As these results indicate, raters unanimously prefer the portfolio rating as a more fair and accurate evaluation of student writing ability. Once again, the raters' self-assessment is not as high, but 75% of those who rate portfolios and essays were convinced they had been more fair and more accurate in rating portfolios. These surveys offer support for the claim that portfolios can be rated reliably.

Frankly, we expected these surveys to support portfolio scoring over the single-essay exam scoring, and they did; however, the degree to which raters endorsed portfolio scoring was remarkable. An analysis of transcripts of the discussions that took place during the calibration sessions suggests that raters were focusing on content and construct validity rather than mere face validity (Stygall).

Rater Training

Our analyses of scoring patterns and raters' own opinions have convinced us that it is possible to score portfolios as reliably as single essays if the training is adequate. We posit several reasons for being able to score portfolios reliably: faculty's initial attitudes toward portfolios, a carefully designed training session, the contents of the portfolios themselves, and the relative ease of reading the portfolios.

A number of participating raters have been using portfolios in their own composition classrooms; some even mentioned that in their survey responses. Our department has a long-standing commitment to composition and a track record of support for innovation. Thus, the population from which raters come is one likely to bring positive attitudes to rating portfolios. What this will confirm to writing program administrators is, of course, that institutionalized, large-scale assessment programs do not exist in a vacuum but are a part of an entire writing program. Still, survey results indicate that even raters who support portfolio rating in principle are less confident in practice about their own ability to score portfolios reliably. A WPA who administers a portfolio scoring session needs to prepare carefully for the portfolio assessment to be successful.

Inexperienced portfolio raters participating in their first portfolio scoring may feel disoriented, no matter how experienced they may be with holistic scoring methods or how positive they may feel about using portfolios in their classrooms. The disorientation is a product of their inexperience in rating portfolios in a large-scale assessment. As experienced holistic raters, they may have a sense of what can be expected in a blue book essay produced within a time limit; and as experienced teachers they may have a sense of what can be expected of their own students in a classroom portfolio; but they understandably ask "What is possible in a multi-sample portfolio produced by students whom I have never taught?" The best way to answer this question is through a thorough training session preceding the scoring.

An effective training, or calibration, session depends upon an effective chief reader and a useful packet of sample or anchor portfolios (see Grogan and Daiker). A rating committee including the chief reader needs to work diligently to select anchor portfolios that illustrate the range of accomplishment of the collected portfolios.[4] In our case, the rating committee consisted of three directors of the Miami Portfolio Program who each read 50-75 portfolios and met for three days of discussion in order to choose anchor portfolios; with experience, the rating committee can now complete the anchor selection in less time. In subsequent years, we have expanded the rating committee to six or seven and have devoted one day to selection. As we will discuss later, there are a number of complications offered by portfolios that the rating committee can anticipate; anchor portfolios can be selected to illustrate any of these complications if the committee wishes to engage the entire group of raters in a focused

discussion of them. Generally, we have tried to assemble a packet of approximately 10-12 portfolios to illustrate our 6-point scoring scale.

Our practice has been to ask raters to read and discuss between 8-10 portfolios, requiring several hours. We begin by focusing on two anchor portfolios at once. Even though this proves time-consuming, as they read through both portfolios, and demanding, as they attempt to remember both portfolios well enough to discuss them afterwards, the paired portfolios begin to create a context for the grading immediately. While assigning a specific score to both portfolios may present raters with difficulty, deciding which of the two portfolios is better does not prove to be that difficult. After the paired portfolios, the training session continues with single portfolios in much the manner of any holistic scoring calibration session. Although the time devoted to the training session is considerable, it seems necessary in order to assure reliable rating. As one rater commented in her survey, "Since this was the first time I have scored portfolios, the long rating session in the morning was extremely helpful, allowing me to understand the differences between scoring whole portfolios and regular timed exams." Another rater commented, "The full morning of training contributed to my accuracy in assigning scores. Discussion of the samples was particularly helpful."

For a number of reasons, then, we are confident at this point that portfolios can be scored as reliably as single-sitting timed essays. The advantages of portfolios (variety, typewritten texts, faculty interest) can work to assure reliable rating, given that a program has carefully laid the groundwork by designing an appropriate portfolio, preparing faculty to accept the program, and training raters. But administrators of portfolio scoring sessions, however confident they can be in the ultimate reliability of their scoring, must be prepared to encounter other problems in rating that are muted or nonexistent in single-essay scoring situations.[5]

Redefining Holistic Rating

Because portfolios are longer than individual essays usually rated in holistic scoring sessions and consist of several pieces of writing, rating "holistically" needs to be redefined. The demands on portfolio readers are great; they must learn to hold their judgment in abeyance not only over the course of a single essay but over the course of an entire portfolio. The challenge increases when portfolios demonstrate unevenness in writing quality or when one piece overshadows all others or when the ideology of the portfolio troubles readers. In the remainder of this essay, we will focus on these difficulties; however, we emphasize that our experience has been that, with practice, raters can learn to rate portfolios holistically.

Just as rating an essay holistically does not mean assigning a score to each paragraph in the piece and averaging those scores to arrive at a final rating, so rating a portfolio holistically ought not mean assigning scores to each piece and then averaging them. The impulse to rate each piece is quite counter-productive. In a portfolio of four pieces rated on a 1-6 scale, for example, a rater who assigns scores of 6 to the first two pieces is not only no longer reading holistically but is also likely to be tempted to score the portfolio as a 6 based only on the first two pieces. This premature assessment subverts the entire point of using portfolios in the first place. Hamp-Lyons and Condon have argued that readers cannot read portfolios holistically if the multiple texts vary from one another in kind. Their point is that readers are bound to consider the multiple texts in light of one another, weighing their strengths and weaknesses, and finally reaching a single judgment based on parts, not on a dominant impression of the whole portfolio. We grant the point that readers inevitably consider the separate parts of the portfolio as they read, but that need not prevent them from assigning a holistically-derived score. In fact, Condon and Hamp-Lyons reach this same and more optimistic conclusion in later research.

We have used the scoring guide to move raters into rating the portfolios holistically. In the general directions, raters should be explicitly reminded not to score individual pieces but rather to withhold judgment until reading all of the written work. As our general directions state, "The portfolios should be read holistically and given a single comprehensive score on a six-point scale ('6' is high and '1' is low). In determining that single score, do not average the four pieces but judge the quality of the portfolio as a whole." We debated whether or not to provide raters with pads for taking notes. The argument in favor was that raters might need help to remember what they have read since each portfolio has many pages and multiple texts. The opposing argument was that the pad might not only slow readers down but also encourage them to score each piece individually. We finally decided not to provide pads, but we did not specifically prohibit taking notes.

Additionally, our scoring guide makes the ability to compose a variety of pieces a central rating criterion. In other words, writers who compose four pieces in some significant way different from one another should score higher than writers who stick to one genre or one strategy throughout their portfolios. For raters to make such judgments, they must read the entire portfolio. Although readers may weigh the different pieces against one another, they should not be scoring them. The description in our scoring guide of a "6" score reads in part: "A portfolio that is *excellent* in overall quality. These portfolios include four distinctive pieces, one from each assigned genre, that excel in several of the following ways. They

demonstrate an ability to handle varied prose tasks with maturity." On the other hand, the scoring guide describes a "1" score as follows: "A portfolio that is *poor* in overall quality. These portfolios include four pieces, but it may be difficult to identify the four assigned genres. There are few or no signs of an ability to handle varied prose tasks competently." In Condon and Hamp-Lyons' study, the ability to handle varied prose tasks was not an explicit criterion for rating. If the chief reader emphasizes the significance of this criterion during the training session, by shaping discussion to focus on whole portfolios rather than on individual pieces, she can encourage raters to rate portfolios holistically. In short, genuine holistic rating is more likely to take place if the rating committee and chief reader foreground its importance and its challenges forthrightly in the scoring guide and in the training session.

The reflective letter that introduces each portfolio may also assist raters in rating holistically. One essay rater's survey reflected on his first experience in holistically scoring essays: "I continue to react really negatively to judging anonymous, decontextualized writing. I don't know what to suggest, but it's an unpleasant experience." Our program's requirement of a reflective letter is in part a response to the problem of decontextualization. Many of the letters create a context for the portfolio by offering raters insights into the rest of the portfolio pieces, thus encouraging raters to withhold final judgments until reading the entire portfolio. For example, our first-year rating committee assigned one portfolio scores of 6, 6, and 4. In the discussion of their scores, one committee member noted that the reflective letter had mentioned all of the remaining pieces. Her explanatory piece, a research paper that explored the topic of organ donation among African Americans, impressed two of the judges, but the one who had mentioned the reflective letter noted, "The main thing I liked about it, I think, is I got the feeling that I was impressed when she talked about it in the letter. What an interesting topic . . . it sounded like she was going to be alive in talking about it and she was." To make a final judgment about the quality of the letter, the rater had to read the explanatory piece as well to see if the letter's claims for it were convincing; additionally, the letter influenced the reading of the explanatory letter later on.

Another student used her reflective letter to provide a context for the remaining papers by indicating which piece mattered most to her. Her portfolio consisted of the letter, a narrative about a newly widowed woman, an explanation of the word "motivation," and an analysis of Saul Bellow's *Henderson, the Rain King*. In terms of length, the Bellow piece was five pages long, while the narrative was only two, and the "Motivation" piece was three pages; however, her letter was very influential in scoring the portfolio by the rating committee. The writer offered one paragraph of background about her narrative and a briefer paragraph about the Bellow piece, the most significant comment being that it is "a condensed version of my senior term paper." Then she included a one-and-one-half page narrative of how she wrote "Motivation," a narrative in which she quoted her instructor's assignment, provided interior monologue of her thoughts, quoted one of her own poems, and reported her dialogues with her instructor about her drafts.

The rating committee discussed her letter and its effect upon their scoring:

> Rater #1: The one thought I did have about this portfolio is that it says in the general directions to give greater weight to the longer and more substantial pieces. While that "Motivation" piece was not long, I would call it substantial, and she certainly had substantial investment in it.
>
> Rater #2: I think she indicated that she thought one of the more substantial pieces was "Motivation" by talking about it in the letter.
>
> Rater #1: True, but all I was saying is that if you go by length, then you should be counting the Bellow piece as the most important one in there, and I didn't get the feeling that she felt that way herself.

It becomes clear in reading this interchange that relationships between the pieces of writing in the portfolios, which surface in both of these reflective letters, encourage holistic rating. In fact, the portfolio begins to seem less an aggregate of separate pieces than a single whole.[6]

The Roller Coaster and Glow Effects

While familiarizing themselves with the portfolios, the rating committee likely will encounter the "roller coaster" and the "glow" effects. The roller coaster effect occurs when pieces in a portfolio fluctuate dramatically in quality; although our experience has been that this effect was not nearly as widespread as we had feared, it nonetheless does occur.[7] The fact that some portfolios do have their ups and downs is another argument for holistic rating. Raters who make their decision too soon may miss the other half of the roller coaster ride that would affect their final rating. Sometimes, however, the roller coaster ride provided by a portfolio proves problematic for raters.

One portfolio evaluated by our rating committee received scores of 6, 5, and 3 from the committee and was thus not chosen to be an anchor; subsequently, it was scored as a 3 during the actual rating session by a first reader and a 5 by a second reader, necessitating a third reading, which produced another 3. Examining a transcript of the rating committee's discussion revealed that the roller coaster effect was probably the cause of all the discrepant readings.

The portfolio began with a letter that two committee members described as "very good" and "great," although the third member called it "ordinary." The narrative that followed—a recounting of how the writer had become separated from her class and subsequently lost during a trip to Athens—was an impressive piece of writing. One committee member described the narrative essay as "highly successful from beginning to end . . . one of the best we're going to see." He singled out for specific praise this passage in which the young woman has finally gotten her bearings: "The only impression that remains is the moment when I entered Omonia Square. The sun was blazing behind the two running figures in the fountain, and miniature rainbows shimmered in the fountain's mist. I could almost hear angels on high singing Handel's 'Hallelujah Chorus,' my relief was so great." A second committee member said she "really liked the voice" in the piece, but the third member felt that the narrative was good, not great, that it was safe and that "she takes no chances" in writing it.

All three committee members agreed that the student's explanatory piece about the Girl Scouts was "utterly pedestrian." "It seemed to have a point but not much of one," said the member who had been most impressed with the narrative. For the portfolio's final piece, entitled "Extract from Captain Stormfield's Visit to Heaven: *Mark Twain's Best*," the consensus was that the writer had relied more on summary than analysis, reserving most of that analysis for the paper's final paragraph.

Essentially, the three committee members agreed that the portfolio reached a peak in the second piece, the narrative, and hit its low point in the third piece, the Girl Scout essay, although the three readers disagreed on the height of the peaks and the depths of the valleys. As the reader who gave the low score of 3 commented, "I was swayed because the last two pieces . . . were really weak."

Because the four pieces of this portfolio were inconsistent in achievement, we want to argue, the portfolio became problematic to score. Although readers agreed that the portfolio was uneven, the up and down of the four pieces led to discrepant ratings by three raters. The issue at stake was the degree to which the portfolio demonstrated the ability to handle varied prose tasks. While we cannot offer a concrete explanation for the discrepant ratings at the actual scoring session, it seems more than likely

that three additional readers were also troubled by the unevenness of the portfolio. Certainly some wildly uneven essays will surface in a single essay holistic assessment, but a portfolio program can be expected to supply more examples of inconsistent work, given the greater volume of writing produced and variety of writing required. The rating committee must be aware of the roller coaster effect, perhaps selecting as an anchor a portfolio that demonstrates unevenness in accomplishment to make certain that this problem is discussed, not ignored.

The glow effect, on the other hand, is more common. An exceptionally strong (or exceptionally weak) piece early in the portfolio tends to shed a glow (or shadow) over the remainder of the portfolio that can affect the overall rating. For instance, one portfolio began with an introductory letter that ended with this paragraph:

> Over the past few years, I've developed new attitudes toward writing, enjoying it rather than dreading it, and viewing each piece not as one completed but as a work-in-progress. There is always a more appropriate word (most often, the one that awakens me out of a sound sleep at 4 a.m. the day after the deadline), a better phrase, room for improvement. I find this stimulating, not frustrating.

Readers agreed that this letter was an excellent one; in fact, in separate scoring, the letter by itself earned the top score of six from two different readers; however, this student's entire portfolio earned only two fives in the portfolio rating session. That the rest of the portfolio dropped off in quality seems clear, but how much did the strength of the letter help the portfolio's overall score? It's hard to know, but it's not hard to surmise that the very strong impression made by the opening letter must have influenced the raters positively.

In the first portfolio discussed earlier, the committee member who had been so strongly impressed by the narrative essay about being lost in Greece commented, "When I read the Greece thing, and I like it so much, does that glow . . . shed its light on the next things that I read? Well, of course, it does. The question is how much. To what extent does it distort your reading?" His score for the portfolio was the highest of the six readers, so perhaps the glow extended for quite some time. With another portfolio read by the committee, the three scores were 5, 5, and 6. The member who scored it highest commented, "I called the analysis a dreary cut and paste . . . but I seemed to have liked her story better than you . . . did. And I know I liked her letter. I give a lot of credit to people who make me chuckle and she did." Once again, a strong early piece cast a strong glow over the rest of the portfolio.

The glow effect also emphasizes the importance of consistency in handling portfolios as they move from rater to rater during the scoring session. Because each piece in the portfolio can, and probably does, color the raters' views of the following pieces, the sequence of the pieces in the portfolio is significant. If raters are to experience the portfolios in the same manner, each rater must read exactly the same portfolio as subsequent raters, so carelessness in handling the portfolios can lead to problems. It is not hard to imagine that the portfolio with the narrative about Greece might have received even more discrepant scores had one rater read the narrative as the final piece in the portfolio and another read it second. We, therefore, staple the four pieces together and number the pages consecutively as a means of ensuring that the sequence of papers remains unchanged during multiple readings.

It is vital to stress to raters that they need to be self-aware, monitoring how they are being influenced by their reading. If readers make a conscious effort to avoid assigning scores until they have read the entire portfolio, they can perhaps avoid being unduly affected by the glow given off by an exceptional piece. Certainly, the rater who is tremendously impressed by the opening piece and skims the remaining pieces is not doing the portfolio justice or scoring holistically. A judicious choice of anchor portfolios can help the chief reader address the glow effect during the training session and discussion.

Politics and Portfolios

Another potential problem exacerbated by the portfolio is political and is a much larger issue facing holistic scoring of any kind. One rater offered this comment in her survey:

> I continue to be concerned . . . by the silence that surrounds critical thought and political awareness in these scoring guides. Consistently, what I hear is 'well-written' even though a piece is critically disastrous. How a responsible teacher of writing (and I think almost everyone in this dept. *is*) can separate critical thought and political subjectivity from writing is unfathomable to me . . . I do think a student whose politics are not well-argued, full of contradictions, and complications in the very institutional precepts against which they pretend to argue, *should* sit in these composition courses, even if they are in touch with their feelings. Room needs to be given to this different pedagogy in the scoring guides as well as in the holistic training discussions we have after scoring.

The issue raised by this comment focuses on the compatibility of holistic scoring and a social constructionist pedagogy, an issue that has been addressed recently (Bizzell), but this essay is not the place to deal with this debate at any length. Our point is to note that the issue is more likely to surface in portfolio rating than in single-essay rating sessions. Undoubtedly, some programs are developing essay prompts today that ask students to address public issues and challenge political assumptions, prompts that will require rating committees to deal directly with the issue of politics and scoring guides. It seems probable as well that some single-sitting essays on not explicitly political prompts may still produce problematic writing of the sort described by the rater. Portfolios by their very nature, however, increase the probability of such problematic writing occurring since each portfolio includes so much more text than a single-sitting assigned essay. How to handle such portfolios is difficult indeed.

Let us illustrate how such problems may arise in portfolio assessment. The rating committee read one portfolio that included a reflective letter; a narrative about a high school football game; an essay entitled "An Ideal Society" that explained the writer's notion of utopia; and an analysis of Robert Frost's "The Road Not Taken." While all three committee members scored the portfolio as either a 1 or 2 on the scale of 6, a problem arose with the explanatory essay. The passage in question reads:

> I think each country's culture is fine the way it is by having their own customs. There will always be social changes and we will have to learn to live with it. We've done good so far.
>
> In each country everyone will treated equal. Just because of your race or religion, you will not be segregated. If you are homosexual or lesbian, you will be segregated and punished. You will be asked to get mental help and teach you that this is morally wrong. We don't need this in our society.

Here is a partial transcript of the rating committee's exchange about this portfolio:

> Rater #1. . . . the homophobia is enough for me to give it a one.
>
> Rater #2. We're going to have trouble reaching anything like rater reliability if people are going to use standards like that.
>
> Rater #1. . . . Regardless of what we say, it seems to me very likely there will be readers who will give it a one when they come across

a passage like that, the argument being I don't want to give institutional approval and credit to a student who thinks that way ... what concerns me here is that if we distribute credits to students that is institutional approval of what they've submitted.

Rater #2. But not, of course, of their ideas.

Rater #1. Well, *we* know that . . . Let's push this case to the limit. Suppose David Duke, the former Klansman . . . the state legislator down south, his kid's coming to Miami and submits a portfolio. The narrative paper is "The Time We Lynched a Nigger." His explanatory paper is "Why Whites Are Superior to Negroes." His lit crit piece of writing is an analysis of one of his father's speeches, and it's all wonderfully well-written. Do you want to give him a six, and give him credit?

Rater #2. No, I don't think so.

Rater #1. . . . I gave this portfolio a one because of my objection to that one passage.

Rater #3 observed later in the discussion that Rater #1 was no longer rating holistically when he based his score on that one passage, a point well taken. Our purpose in this section of our essay, however, is to note that such problematic situations are going to arise more often in portfolio assessment. It is important not to magnify this problem since it did not present a major obstacle to our rating. Still, highly significant issues are raised by portfolios such as the one discussed above, and the rating committee should think through how to handle such situations if and when they arise. We decided to advise our raters to send back to the chief reader any portfolios that they found they could not rate. This "solution," however, is not entirely satisfactory because it begs the question of what to do with such portfolios in the first place, since the chief reader simply sends the portfolio back out to be rated until someone rates it or else rates it her/himself. It also assumes that raters would be too troubled to rate the portfolio, a naive assumption given that the member of the rating committee most troubled by the portfolio did not hesitate to score the portfolio lower because of the disturbing passage.

The problem raised here is not really with portfolio assessment but with all holistic assessment. Holistic rating, like all other aspects of the composition curriculum, should be subject to intense examination and analysis, and portfolio assessment is very likely to intensify this scrutiny.

The politics of assessment promises to be a site of continued debate—and rightly so with portfolio assessment no doubt affected by that debate—but we ought not mistakenly see portfolios as the cause of the problems.

Conclusion

Our purpose in this essay has been to advance an important claim and continue an important ongoing discussion. Our claim, which we have attempted to substantiate in the first section of the article, is that portfolios can be reliably rated in a large-scale assessment situation, given proper preparation by a writing program administrator. More research, of course, needs to be conducted to determine whether the reliability we reached is reproducible and how best to make sure portfolio rating remains reliable.

In the second half of the essay, we added to the growing discussion of portfolio assessment by sharing our experiences with other writing program administrators. Portfolio assessment is similar enough to other holistic assessment methodology that some of the problems it presents are not new but simply exacerbated by the longer, more complex group of writings presented in portfolios. On the other hand, portfolio assessment is different in important ways from the kinds of holistic writing assessment with which many of us are more experienced, therefore presenting different problems that need to be addressed if reliable rating is to be achieved.

Portfolios are not the panacea to our assessment needs, as Edward White has recently pointed out (*New Directions*), and they are not simple to administer or rate. To a great extent, the future of portfolio assessment rests on the experiences, observations, and insights of writing program administrators in the midst of administering portfolio programs. Complex and complicated as they are, portfolios remain our best tool for assessing student writing ability. Our purpose has been to focus attention on the actual machinery of scoring portfolios. We have outlined some practical methods of gaining reasonable reliability, but we are aware that serious concerns about validity and reliability in scoring portfolios remain. Portfolios have clearly arrived; we hope that this article points the direction for continued conversation about portfolios as an assessment method.

Notes

1. Our research was funded in part by a grant from the National Council of Writing Program Administrators (WPA).

2. Hamp-Lyons and Condon (CCC) present data that show their raters reaching scoring decisions after reading a single page of a portfolio, which would, of course, account for high reliability in a less-than-welcome way. In a later piece (Condon and Hamp-Lyons, *New Directions*), they report having found a way to retrain readers. See p. 17 of our article for further discussion.

3. Of course, bringing the personal into the scoring can be problematic at times. See Laurel Black, Donald A. Daiker, Jeffrey Sommers, and Gail Stygall, "Writing like a Woman and Being Rewarded for It? Gender, Assessment, and Reflective Letters from Miami University's Student Portfolios" for discussion of the effects on scoring the reflective letters.

We also scored reflective letters separately. If we compare the scores received by the individual letters to those received by the entire portfolios, we find that only 53% of the letters were scored within one point of the portfolio score.

We conclude that our reliable scores on portfolios are not caused by premature judgments based solely on the reflective letters. Clearly, portfolio raters are reading the entire portfolios before determining a score.

4. David W. Smit, "A WPA's Nightmare: Reflections on Using Portfolios as a Course Exit Exam," argues that selecting anchor portfolios is virtually impossible given the perplexing variety of portfolios. Careful construction of the guidelines to students, however, can provide concrete descriptions of the required writing, thus preventing variety in portfolios from descending into chaos. For a full description of Miami's guidelines to students, see Black et al., *Handbook on Writing Portfolio Assessment: A Program in College Placement.*

5. The rating committee must decide how to deal with an incomplete portfolio and a portfolio that provides more than one piece in a required genre. While it is certainly possible that an incomplete essay might be produced during a single-sitting essay examination, the situation does not occur often enough to be a major concern. Certainly, there should not be a major problem with incomplete portfolios, but the odds increase when requirements made of students increase. The incomplete portfolio can either be deemed unscorable as a non-responsive essay would be, or it can be scored on the 1-6 scale, with raters taking into account that the portfolio is incomplete. The Miami program chose to set aside incomplete portfolios, weeding them out during the preparation phase; these portfolios were never distributed to raters, and they received scores of 0, meaning "Not Rated." Portfolios that included two narratives, instead of a narrative and a textual analysis, were scored, with the portfolios receiving lower scores because of their inability to meet all of the scoring criteria.

6. The reflective letters not only affect the rating of the portfolio in important ways, but they also present a challenge to rating committees accustomed to the holistic scoring of single essays. Unlike the essay-scoring situation in which raters merely read the essays, portfolio-scoring situations provide raters with a meta-commentary by the student about the written work. Reading the letters is a new task and one for which the rating committee must be prepared. Our ongoing research has been focusing on these letters, which are fascinating pieces of writing; part of our research has been an effort to develop a scoring scale for the letters themselves to see if they can be rated reliably as separate pieces. We have, in fact, rated the letters; only 14 of 270 letters required a third rating. It is thus reasonable to expect raters to grow accustomed to rating a "new" kind of piece in large-scale assessment situations, given enough practice.

7. Condon and Hamp-Lyons (*New Directions*) describe this phenomenon without naming it in reading protocols written by their raters. One rater writes, "After I read the first essay, I was sure this would be a Practicum placement [the lowest score], but the impromptu and the second revised essay changed my mind."

Works Cited

Belanoff, Pat, and Marcia Dickson. "Introduction," *Portfolios: Process and Product*. Eds. Pat Belanoff and Marcia Dickson. Portsmouth, NH: Heinemann, 1991. xix-xxiv.

Belanoff, Pat, and Peter Elbow. "Using Portfolios to Increase Collaboration and Community in a Writing Program." *Portfolios: Process and Product*. Eds. Pat Belanoff and Marcia Dickson. Portsmouth, NH: Heinemann, 1991. 17-29.

Bizzell, Patricia. "What Can We Know, What Must We Do, What May We Hope: Writing Assessment." *College English* 49 (1987): 575-584.

Black, Laurel, Donald A. Daiker, Jeffrey Sommers, and Gail Stygall. "Writing like a Woman and Being Rewarded for It? Gender, Assessment, and Reflective Letters from Miami University's Student Portfolios." *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

———. *Handbook of Writing Portfolio Assessment: A Program for College Placement*. Oxford, OH, 1992.

Camp, Roberta. "The Place of Portfolios in Our Changing Views of Writing Assessment." *Construction vs. Choice in Cognitive Measurement*. Eds. R. Bennett and W. Ward. Hillsdale: Lawrence Erlbaum Associates, forthcoming.

Charney, Davida. "The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview." *Research in the Teaching of Writing* 18.1 (1984): 65-81.

Condon, William, and Liz Hamp-Lyons. "Introducing a Portfolio-Based Writing Assessment: Progress through Problems." *Portfolios: Process and Product*. Eds. Pat Belanoff and Marcia Dickson. Portsmouth, NH: Heinemann, 1991. 231-247.

_____. "Maintaining a Portfolio-Based Writing Assessment: Research That Informs Program Development." *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

Elbow, Peter. "Foreword." *Portfolios: Process and Product*. Eds. Pat Belanoff and Marcia Dickson. Portsmouth, NH: Heinemann, 1991. ix-xvi.

_____. "Will the Virtue of Portfolios Blind Us to Their Dangers?" *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

Ford, James E., and Gregory Larkin. "The Portfolio System: An End to Backsliding Writing Standards." *College English* 39 (1978): 950-955.

Grogan, Nedra, and Donald A. Daiker. "Team-Grading in College Composition." *WPA: Writing Program Administration* 13.1-2 (1989): 25-33.

Hamp-Lyons, Liz, and William Condon. "Readers' Responses to Portfolios." *College Composition and Communication*. Forthcoming.

Huot, Brian. "The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends." *Review of Education Research* 60.2 (1990): 237-263.

_____. "Beyond the Classroom: Using Portfolios to Assess Writing." *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

Smit, David W. "A WPA's Nightmare: Reflections on Using Portfolios as a Course Exit Exam." *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

Stygall, Gail. "Comparing Validity in Writing Assessment: Impromptu Writing Exams vs. Portfolio Assessment." Conference on College Composition and Communication. Boston, March 1991.

Tirrell, Mary Kay. "A Natural Choice for Evaluation: Links Between Portfolio Assessment and Administrative Style." New Directions in Portfolio Assessment Conference. Oxford, Ohio, 1992.

White, Edward M. "Portfolios as an Assessment Concept." *New Directions in Portfolio Assessment*. Eds. Donald A. Daiker, Jeffrey Sommers, Gail Stygall, and Laurel Black. Portsmouth, NH: Heinemann, Boynton/Cook, forthcoming.

Yancey, Kathleen Blake. "Teachers' Stories: Notes toward a Portfolio Pedagogy." *Portfolios in the Writing Classroom: An Introduction*. Ed. Kathleen Blake Yancey. Urbana: NCTE, 1992. 12-19.