Selecting and Using Sample Papers in Holistic Evaluation

DONALD DAIKER AND NEDRA GROGAN

Anyone experienced in holistic scoring understands the vital importance of sample papers. By general consensus they are, next to the chief reader, the single most important element in a successful scoring session. But if test developers or chief readers look to scholarship for help in selecting and using sample papers, they will not find much. In the literature on writing assessment in general and holistic evaluation specifically, the selection and use of sample papers is an issue lost among larger concerns of reliability and validity (Charney) or among more “obvious” concerns like writing prompts (Ruth and Murphy).

The variety of names for sample papers—anchor papers, prototypes, range-finders—suggestions their multiple characteristics, manifold functions, and problematic nature. A review of the literature shows that sample papers are supposed to fit all of the following descriptions: papers with attributes that characterize each score level and exemplify various point scores (Cooper 41); papers which have “the desired normative distribution” (Miles Myers 32); papers that are representative, especially of the middle range of a single score (Northwest 65); papers “which illustrate the full range of performance on the assignment” (Odell and Cooper 36); papers that “virtually all readers” will agree upon (Spanel and Stiggins 20); and papers that illustrate “the distinctions in quality described by the scoring guide” (White, Teaching 160). While these are overlapping descriptions, the differences among them complicate the role of sample papers while indicating little about selecting and using them.

Assessment literature often sidesteps the troubling nature of sample papers with a common refrain: “It all depends on the particular testing situation.” Despite the appropriateness of this qualification (and its usefulness will become evident as we continue), it cannot serve as the answer to all the critical questions which precede a reading. Our purpose here is to outline the series of decisions involved in selecting and then using sample papers. We want, in particular, to make test-developers aware that they have choices to make and that their choices will have serious consequences for student and evaluator alike.
Selecting Sample Papers

Although sample papers have come to seem an integral part of holistic evaluation, the two did not develop simultaneously. Holistic evaluation came first, followed by sample papers only after researchers found readers to be unreliable without them. Coffman and Kurfman were among the first to recognize the "idiosyncratic fluctuations" of readers in both analytic and holistic scoring situations. They conclude—and their recommendation has been embraced by the Educational Testing Service and widely accepted by the profession—that sample papers must be used regularly and repeatedly to train readers in holistic evaluation.

The procedure of selecting sample papers to train readers begins with relatively uncomplicated matters of logistics and planning. Because results are often needed immediately, especially for placement and proficiency examinations, the selection session is commonly held as soon after the writing as the selection committee members can assemble. The selection session should take place in any quiet setting with a table large enough to spread out stacks of papers.

Selecting sample papers requires more than a single person. Chief readers invite trouble (and perhaps even disaster) by selecting sample papers on their own. At some point during the subsequent training session, "independent" chief readers can be sure that their evaluations of sample papers will be challenged and that they will have to explain and defend their scores (Daiker). An explanation will be more persuasive if the score represents the consensus of the full selection committee and if committee members provide supporting comments of their own. In our experience, a five-member selection committee is ideal: three is the bare minimum, while more than seven becomes unwieldy. Also important is the choice of committee members, and here common sense prevails: taking into account local political realities, it is best to choose members who have experience in holistic evaluation, who have earned the professional respect of their colleagues, and who do not define compromise as loss of face.

The ultimate goal of sample paper selection is reader consistency. The immediate goal is the identification of a set of papers to be used in training readers to reach consistency. How many papers are needed depends to some extent on paper length (shorter papers take up less training time) but more importantly on the range of the scoring scale and the number of reading days. More papers are needed to illustrate the nine-point scoring scale used by the Educational Testing Service to score Advanced Placement essays than for the four-point scale used by several universities for placement purposes. By the same token, ETS requires many more samples for its six-day reading than Miami University needs for a reading that takes only half a day. With the popular six-point scale, two papers for each point (a total of 12) are usually more than enough for a single day of reading. For each additional day, select another paper for each point of the scale.
To obtain the twelve or eighteen sample papers, committee members must read papers from the batch to be evaluated. If papers have been randomized, each member grabs forty to fifty papers and retreats to a quiet spot. If the papers have not been randomized, and especially if they are arranged by class, school, or ability level, the chief reader provides each member with a roughly representative set of papers—papers that can be expected to approximate the range of the entire population.

With a set of papers in hand, committee members start reading on their own. And that's all they do: read. They temporarily forget about scores and scales because the primary purpose of this initial reading is to get a general sense of how students have responded to the assignment: to begin observing differences between more and less successful papers, and to begin developing a feel for the range of responses. It is important that these first papers not be evaluated or placed in piles corresponding to the points of the scoring scale. Since holistic scoring is nothing more than the rank ordering of papers from one particular testing situation, scoring cannot begin (and shouldn't be attempted) until a number of papers from that testing situation have been read.

Only after an hour or more of reading without scoring should readers move to the second step in the selection process: the preliminary scoring of papers. Readers should put aside the papers they've already read (they've served their purpose) and take new ones. Now is the time to begin applying the scoring scale deliberately and thoughtfully. The steps are simple. Don't write on the papers. Instead, place each into a pile according to the score you've assigned it; for a six-point scale there will be six piles. When there are two or three papers in each pile, it's time to review your scoring. Start at either the upper or lower end of the scale and reread, in order, the papers in each pile, making any necessary adjustments as you go. The objective is to choose, for each point of the scale, one or two papers to recommend as sample papers to the selection committee.

When the selection committee meets, members bring the papers they are proposing as samples. The chief reader begins the formal part of the meeting by asking everyone to place a single, unmarked paper in the middle of the table. The chief reader makes clear (and this is a crucial point) that each member is free to choose a paper from any point on the scoring scale. The more points represented, the better. Problems lie in the other direction—when the chief reader invites members to submit, say, samples of upper-range papers. Such an invitation may speed up the selection process, but it can only bias the results. As Paul Diederich warns, "Grading is such a suggestible process that we find what we expect to find" (12). If we are told to expect upper-range papers, we'll find them. The best way to limit bias is not to create expectations.

Each member then takes a proposed sample paper, reads and scores it, records the score and paper number on a separate slip of paper, and then
repeats the process until all the papers are scored. Scoring slips are now passed to a designated committee member, who records the scores on a chalkboard for all consider. It's a good idea for the chief reader to begin discussion by focusing on the paper that comes closest to gaining a consensus score. If one of the proposed samples receives, say, a 3 from everyone, that's the paper to start with. The chief reader announces the scores and then asks, "Why did you score this a 3?" The responses will be useful in several ways: they provide the chief reader (who is taking notes) with material for the subsequent training session; they begin the process leading to a written statement of criteria (sometimes called a rubric); and they make it easier to select sample papers representing other points on the scoring scale. Unless the responses unexpectedly reveal confusion or disagreement, the paper that everyone graded a 3 is then set aside as one of the samples to be used during the training session.

It makes good sense to move, in order, from the papers that have produced the most initial agreement to papers that have produced the least. Of course, a paper need not have received exactly the same initial score from all members to qualify eventually as a sample paper. But cautious chief readers will not select as a sample paper one that has received more than a single divergent score or one whose single divergent score is off by more than a point. Nor will they choose a paper with a single, one-point divergent score unless the divergent scorer willingly changes his or her mind as a result of discussion.

From the first-half dozen papers proposed as samples, at least one usually gains consensus and is accepted. But the proposed samples that fail to gain consensus are essential to the selection process. These papers need to be discussed fully in order to lay the basis for future agreement. The lack of consensus is a sure sign that committee members are not yet employing similar standards and criteria; discussion of specific papers is the most tactful and effective way of making members aware of the situation and of encouraging them to come together. Not that members are always willing to change an evaluation of a paper: it is not easy to admit to being "off" or to "missing one." Indeed, as we later suggest, it is sometimes not an instance of "missing one" at all but of bringing new perspectives to a complex task. Still, when members recognize that their evaluation varies significantly from that of colleagues they respect, they silently make the kinds of adjustments that lead to future agreement.

The process of submitting, reading, scoring, and discussing papers continues until through consensus two or more sample papers have been selected for each point of the scoring scale. Then the committee must decide how many sample papers at a time should be given to readers during the training session, in what sequence the papers should be presented, and how the scoring criteria should be described in the rubric given to readers.

How many at a time depends in large part on paper length. ETS training
sessions for readers of AP examinations in English typically begin with four papers: a 2, a 4, a 6, and an 8 on the nine-point scale. But AP essays are written in thirty-five or forty minutes and tend to be short, averaging just over 200 words. Even so, subsequent batches of papers at AP training sessions usually consist of two or three papers, never more than four. With longer papers, especially those in the range of 500 words, it seems more reasonable to use pairs of papers from the beginning.

The central principle in sequencing papers is to guide readers into making increasingly more sensitive scoring decisions. For this reason, it makes sense to start with a pair of papers that present a clear contrast in quality without either paper representing the very top or bottom of the scale. With the six-point scale, we like to begin with a 5 and a 2, or a 5 and a 3. Either pair, but especially the 5 and 2, is likely to produce consensus, at least in terms of which is the better of the two papers. It also makes sense to begin with papers that represent the middle of both the top range (4-6) and bottom range (1-3). Once readers experience mid-range papers, they can more readily identify papers at the ends of the scale.

If the initial pair is a 5 and 2, the committee might choose a 6 and a 3 or a 4 and a 1 as its second pair. In either case, it makes sense to choose papers that have earned a “solid” score until each point on the scale has been represented once. A solid 5 or a solid 2 is one that committee members agree occupies the middle of the 5 or 2 range (Northwest 65). As this statement implies, each point on any scoring scale represents a range of scores; that’s why teachers are fond of adding a plus or minus to the letter grades they assign. Once the committee chooses “solid” papers to illustrate every point on the scale, it decides whether to include papers judged to occupy the bottom or top of its point range (a high 3, a low 5). A final decision is whether to use borderline papers (a 3/4 split; a 5/6 split) in the training session that follows.

After sample papers have been placed in order of use, labeled, photocopied, and bound in a packet for the upcoming training session, leaders must devise a rubric, a statement of scoring criteria. The rubric characterizes papers representing each point of the scoring scale (see the sample rubric below). Creating a rubric must be the last step of sample paper selection, for until papers have been chosen and discussion completed, no scoring criteria exist.

Without the advantage of the discussion that accompanies the selection of sample papers, a rubric will be imposed rather than “inferred,” and most researchers agree that scoring criteria must be allowed to develop inductively from a particular set of papers (Miles Myers; Odell and Cooper). Because of the very nature of holistic scoring—a procedure limited only to a rank ordering of papers from a single testing situation—the scoring criteria must evolve “naturally” from the testing situation and from committee members’ dialogue about each score level. Experienced readers’ choices of represen-
tative writing samples become the foundation of the scoring guide.

Sometimes, of course, chief readers are not able to make use of the full selection process described here. Constraints of time and money may force procedural compromises. The least damaging compromise, in our view, is simply to eliminate the first two steps above: the chief reader, and not the committee members, reads widely among the essays and makes a preliminary selection of sample papers. These papers are then blindly scored by the committee members and the chief reader's preliminary scoring is either accepted or rejected, paper by paper. The most damaging compromise is for the chief reader to select samples on his or her own or to bias the reading of the selection committee with directions such as, "See if you don't agree that this is a solid 4 paper." Compromises are more difficult to make at the next stage of the holistic procedure: the use of sample papers.

Using Sample Papers

Despite its complexities, the selection of sample papers is only preliminary to their use in a training session. Since the goal of training is to guide readers toward consistency in evaluation, the chief reader tries to use sample papers in such a way as to achieve reader agreement (or calibration) quickly and harmoniously. Even before the first sample paper is distributed, the chief reader and committee members will have opportunities for creating an environment conducive to consensus. The choice of a well-lighted and comfortable room, the availability of snacks and beverages, a warm welcome and personal introductions, and perhaps even reader involvement in preliminary tasks of sorting and distributing—all these contribute to the sense of collegiality that makes agreement possible. Just as important (because it builds confidence in both the leaders and the procedures) is a sense of professionalism that can be established through seemingly insignificant details like name cards at assigned seats, designated reader numbers, coded papers, and a printed schedule of the day's activities.

The choice of a room for the reading requires more planning than a call to the room-reservations office. Leaders will want to decide in advance between tables and theater seating. If tables are used, as ETS does, committee members commonly become table leaders, thereby assuming primary responsibility for using sample papers to train and calibrate the five-to-ten readers seated at their tables. But for small holistic sessions, theater seating is more efficient, more economical, and perhaps equally effective. With theater seating, the chief reader stands at the front of the room with the readers seated in a semi-circle facing the chalkboard or overhead screen.

Once the setting has been established and a collegial atmosphere created, the chief reader begins formal training by first summarizing the distinguishing features of holistic evaluation (White Teaching 24-26) and then distributing copies of the writing assignment or prompt. Here is a sample assignment:
**College Composition Proficiency Examination**

**Instructions**

Spend about 20-30 minutes reading and thinking about the poem “Next Day” by Randall Jarrell. During this time, you may want to free write, take notes, construct an outline, or write a partial draft. When you feel you’ve come to an understanding of the poem, write an essay of at least 500 words explaining how the poetry works to create your impression of the speaker. You may wish to evaluate this character in light of your own experiences or in relation to characters you know from movies, television, or other works of literature. You may also consider how this character reflects the poet’s values or attitudes. Whatever your views of the character, they should grow out of a careful examination of the text, including features such as voice, figurative language, and sentence structure.

**Criteria for Evaluation**

The instructors of composition and literature who evaluate your essay will be looking for evidence of your ability to read intelligently and insightfully and to write clearly and persuasively. In particular, they will be looking for:

- A unified, well-organized analysis offering a persuasive interpretation of the poem
- An essay rich in detail, one whose key assertions are supported by specific references to the text
- A well-written essay, demonstrating effective word choice, sentence structure, paragraph construction, and control of other writing strategies

[Text of poem follows]

When the writing assignment involves analysis and interpretation of a piece of extended writing—an essay, a short story, a long poem—it’s sound practice to provide readers with the text at least a day in advance of the reading. But even when readers receive the text beforehand, it makes sense to allow several minutes for rereading and reviewing. The chief reader then invites questions about the instructions and, sometimes, about the text itself. Although it is important for the chief reader to make sure readers understand what the writing assignment asks of students, a discussion of the text’s meaning at this point will tend to shift the focus to interpretation and away from evaluation. So it may be wiser to encourage only questions that relate to the instructions or the testing situation.

The chief reader’s last task before turning to sample papers is to define
the rubric. Each reader should receive a copy of the rubric so that it can be referred to both during the training session and during the scoring of papers. Here is a sample statement of rating criteria that a chief reader and committee members designed after they had selected sample papers but before the training session:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>The best papers take charge of the assignment, offering a range of approaches to the poem. They are generally longer, including numerous interpretive insights and often showing the writer's ability to incorporate materials from outside the poem into a coherent essay. They are stylistically felicitous, offering satisfying turns of phrase or interesting sentence structures.</td>
</tr>
<tr>
<td>5</td>
<td>These writers likewise offer interesting and insightful interpretations of the poem, but they may be less sophisticated in their handling of organization or theme. Like the 6 paper, this level includes abundant supporting references to the poem without digression or extended summary. The style is smooth, sometimes artful.</td>
</tr>
<tr>
<td>4</td>
<td>This rank represents the better performance in the middle range, but these papers fall below the best because of their unevenness. The writer may be less sensitive to the character; personal experience or reference to other texts may not work well with the interpretive paragraphs. Or the paper may lack balance: the reading may be impressive but the organization too predictable. But these writers will still offer a sufficient number of references to the text. While these papers are generally mechanically correct, the style can be less than elegant.</td>
</tr>
<tr>
<td>3</td>
<td>At this level, the papers are clearly unsuccessful, both as interpretations and as compositions. The interpretation of the poem will often read like a summary with few specific references to the text. Use of conventional literary critical tools may seem labored. Organization may be flawed or predictable. Style may be overly simple or marred by odd diction.</td>
</tr>
<tr>
<td>2/1</td>
<td>At the lowest level, students usually have much less to say. Content is marked by repetition of empty phrases. The writer may rely heavily on summary. Organization may be tediously formulaic or random. While few of the papers will contain glaring or numerous mechanical errors, most will show a lack of control of sentence structures and a limited vocabulary.</td>
</tr>
</tbody>
</table>

The chief reader reads the rating criteria aloud and then turns to the heart of the training session: the distribution and scoring of sample papers.
For each sample, readers are asked to read quickly and holistically, attending to the effect of the whole rather than to any individual part. Once readers arrive at a score, they are required to write it down; that way, they will be less likely to report a different score when the chief reader collects votes. Whether scores are collected by a show of hands or by slips of paper, they should be recorded on a chalkboard or overhead projector so that readers can sense the level of agreement (or disagreement) on each sample. When readers see, for example, that all their scores landed in the lower range, they also see evidence of shared judgments and common goals. This “success” in judging a paper gives the reading an important atmosphere of community, and from this initial consensus, the community may paradoxically be able to accommodate later disagreement and even conflict.

Finally, after all the logistical decisions have been made, the chief reader can distribute the samples and start the training session. Ideally, the samples have been photocopied, alphabetically labeled in tentative order of use, and stapled together as a single packet, one for each reader. Handing out packets rather than individual papers saves distribution time and keeps aides from distracting trips around the room. The chief reader can begin, then, by asking the group to read and score papers A and B in their packets.

Whether the first pair of sample papers consists of a 5 and 2, or any other combination, the chief reader takes into account the level of rater agreement before moving to a second pair. With acceptable agreement on the first pair (almost all scores within a single point of the agreed-upon score), the training session can move comfortably to sample papers representing other points on the scoring scale. But if the first pair of papers receives a wide range of scores, the chief reader may want to choose a second pair of papers representing the same scores as the first. The choice of which pair to use next may also depend on the context of the testing situation. For example, if it seems clear that the distinction between a 4 and a 5 is crucial, separating students who will receive credit and/or advanced placement from those who will not, it makes more sense that the second pair consist of a 6 and a 4 than a 3 and a 1.

The most difficult situation is substantial rater divergence from scores earlier established by the leaders. This is the moment every chief reader dreads because the success of the session rests on a decision that has to be made quickly. Although wise chief readers encourage discussion to give themselves time for thought, they must eventually decide whether to insist on the score chosen by the leaders or to accept the score favored by the raters. In some instances, the decision is obvious: raters are often reluctant to assign the top score of 6, especially early in a training session, and they need only to be assured by leaders who have read hundreds of papers that such papers are uncommon. Precisely this situation arose at a recent Miami University scoring session: a sample paper chosen as a 6 by the leaders was rated a 6 by thirty raters, 5 by another thirty, and 4 by four others. Because rater disagreement occurred early in the training session and because the leaders...
had previously agreed that the paper was not only a 6 but a high 6, the chief reader knew immediately that she could not compromise with the raters and declare the paper a 5/6 split. Had she done so, the scoring range would have narrowed considerably and rare 6 papers would have become non-existent. So the chief reader gently but firmly tried to convince dissenting readers that the paper deserved the score of 6. She first asked for comments from raters who had given it a 6. They responded enthusiastically, characterizing the paper as “striking,” “insightful,” and “powerful.” In fact, one of the department’s most experienced and respected members said he thought the paper would be among the very best he would read all day. Heads nodded in agreement. When the chief reader then asked for comments from raters who had scored the paper a 5, their responses sounded unconvincing. After the reminder that student writers are to be rewarded for what they do well (rather than penalized for their weaknesses), the chief reader ended discussion by announcing that the leaders (who had together read over 200 papers) were unanimous in scoring the paper a strong 6.

It was an entirely different situation several sample papers later, however, when an essay intended to represent the score of 5 received a 5 only from the leaders who had chosen it; the other raters divided evenly between 3 and 4. For the chief reader, the discrepancy between scores assigned by the leaders and raters presented a delicate rhetorical problem. One of her options was to ignore the leaders’ score of 5; after all, readers were not aware of the discrepancy—only the leaders were. Another option was to announce the difference immediately and try to defend the paper as a 5. Instead, since the chief reader had carefully reread the paper and found herself no longer convinced that it merited the score of 5, she allowed discussion to center on the paper as a 3 or 4 and only afterwards admitted that the leaders had scored it a 5. When a loud buzz arose among the raters, the chief reader laughed and modestly acknowledged, “But we blew it. You guys are good!”

Thus a chief reader may on occasion want to yield to the judgment of raters. Not to yield when the vast majority of raters disagree with the leaders’ score late in a training session is to risk chaos. It seems wiser, under such circumstances, to assume that raters are now better calibrated than the leaders, and to move quickly to another sample paper.

Question of Authority and Consensus

As the Miami University scoring session makes clear, the most difficult questions concerning sample papers in holistic evaluation involve issues of authority and agreement. What may be particularly disturbing about holistic scoring is the need for consensus in a field that privileges dialectic and is based on academic freedom. In his important and provocative “Reality, Consensus, and Reform in the Rhetoric of Composition Teaching,” Greg Myers warns of dangers inherent in the “construction of a sense of general agreement”—the central aim of any holistic scoring session. Myers correctly
Samples in Holistic Evaluation 169

points out that achieving consensus always means the suppression or exclusion of some point of view. Within the composition classroom, where it has become standard practice for small groups of students to collaborate on and critique each other’s writing, Myers is especially troubled by the drive toward general agreement. He fears “that the teacher has merely embodied his or her authority in the more effective guise of class consensus. This guided consensus has a power over individual students that a teacher cannot have alone” (159). Although Myers does not directly address issues in writing assessment, his essay implicitly challenges any system of evaluation in which authority cloaks itself as consensus. The idea of consensus comes under further attack in Joseph Harris’ “The Idea of Community in the Study of Writing.” For Harris, consensus is not only impossible to achieve but undesirable: every community must experience a certain amount of conflict and struggle if it is to change and grow. Harris goes on to define a community (especially an academic community) as one in which “competing beliefs and practices intersect with and confront one another.”

These warnings against consensus and the misuse of authority, however, seem to have only limited application to holistic scoring, largely because the group that ordinarily gathers for a holistic scoring session differs in fundamental ways from the group that constitutes a classroom. Indeed, holistic scorers generally meet Harris’ definition of an “academic discourse community”: “a free and voluntary gathering of individuals with shared goals and interests—of persons who have not so much been forced together as have chosen to associate with one another” (15). By contrast, students come to their composition courses because they are fulfilling degree requirements at a time and place specified by the registrar’s office.

The matter of authority is a second fundamental difference between the scoring session and the classroom. Authority in the classroom clearly resides with the instructor; efforts to share authority—by, say, allowing students to help make decisions about course assignments or methods of assessment—do not fundamentally alter the balance of power or change the fact that one person in the classroom is felt to be superior in most ways that count. The situation is quite different with holistic scoring, where all participants are professionals who are more likely to be equals than subordinates. In the Miami University scoring session, an untenured assistant professor served as chief reader for colleagues that included tenured associate and full professors. During an Advanced Placement scoring session run by ETS, an unpublished high school teacher served as Table Leader for several college professors, one the author of an important composition textbook and the other a leading Steinbeck scholar. In holistic scoring, then, authority is likely to be more limited in scope (to a specific assignment, rubric, and set of papers) and in time (to a morning, a day, or at most a week).

Not only is authority less potent in a holistic scoring session than in the classroom, but consensus is more desirable. Myers is right in warning that in
the classroom, "the tendency to unthinking conformity is always there when consensus is used to enforce standards" (159). Such consensus may indeed mean "eliminating or at least concealing diversity and conflict" (160). But in holistic evaluation, consensus cannot be defined as dangerous if we are committed to the direct assessment of writing. After all, without general agreement or "calibration" among raters, holistic scoring becomes an unreliable means of assessment. Who could then blame test-makers for reverting to multiple-choice and short-answer tests which, however questionable their validity as measures of writing competence, can be demonstrated to be remarkably reliable?

In selecting and using sample papers, chief readers will therefore want to take into account both the need for general agreement and the value of disagreement. They will encourage consensus without burying conflict, guided by the distinction Edward M. White posits between objective reality on the one hand and more complex phenomena on the other. If we are measuring objective reality (for example, the weight of a backpack or the number of lines in "Fern Hill") differing results signal an error of some kind. But if we are evaluating a writing sample or a work of art, where there is no one right answer, a difference is not only not an error but "positively valuable" in helping us "see the work in question more clearly and estimate its value more intelligently than simple unanimity would" (Developing 94). After all, most of us would agree that questions like "Which is Virginia Woolf's best novel?" and "Are Rembrandt's self-portraits more powerful than Van Gogh's?" do not have a single correct answer.

Without objective reality as a guide, chief readers must decide for themselves when to wield and when to yield authority. In the Miami University training session, the chief reader was willing to acknowledge conflict, let the discussion play out, and then compromise with raters. With one paper she returned to the scoring community the authority she had taken away with another—perhaps because she realized that the goal of holistic scoring is not consensus between the leaders and raters but consensus among the raters themselves. Whether it was her logic or her intuition that saved the collegial atmosphere of the reading is hard to tell. But the outcome of the conflict was a happy compromise: the authority figure and the community of readers took turns making the scoring decisions.¹

¹The writing of this article was made possible by a grant from the Committee on Faculty Research of Miami University.
Samples in Holistic Evaluation 171

Works Cited


