Collaborative holistic evaluation: theoretical underpinnings and classroom use

Grant Moser
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Educational Assessment, Evaluation, and Research Commons, and the Rhetoric and Composition Commons

Recommended Citation
Moser, Grant, "Collaborative holistic evaluation: theoretical underpinnings and classroom use" (1992). Retrospective Theses and Dissertations. 22.
https://lib.dr.iastate.edu/rtd/22

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Collaborative holistic evaluation: Theoretical underpinnings and classroom use

by

Grant Moser

A Thesis Submitted to the Graduate Faculty in Partial Fulfillment of the Requirements for the Degree of

MASTER OF ARTS

Department: English
Major: English (Rhetoric and Composition)

Iowa State University
Ames, Iowa
1992
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>CHAPTER ONE: GRADES AND PROBLEMS</td>
<td>5</td>
</tr>
<tr>
<td>Grades</td>
<td>7</td>
</tr>
<tr>
<td>Problems</td>
<td>15</td>
</tr>
<tr>
<td>Validity</td>
<td>15</td>
</tr>
<tr>
<td>Reliability</td>
<td>22</td>
</tr>
<tr>
<td>Topic Bias</td>
<td>25</td>
</tr>
<tr>
<td>Summary</td>
<td>29</td>
</tr>
<tr>
<td>CHAPTER TWO: THEORETICAL UNDERPINNINGS</td>
<td>31</td>
</tr>
<tr>
<td>Collaboration</td>
<td>31</td>
</tr>
<tr>
<td>Holisticism</td>
<td>35</td>
</tr>
<tr>
<td>Summary</td>
<td>39</td>
</tr>
<tr>
<td>CHAPTER THREE: COLLABORATIVE HOLISTIC EVALUATION</td>
<td>41</td>
</tr>
<tr>
<td>Assignment Design</td>
<td>42</td>
</tr>
<tr>
<td>Evaluation Methodology</td>
<td>44</td>
</tr>
<tr>
<td>CHE and Three Problems in Grading</td>
<td>49</td>
</tr>
<tr>
<td>Validity</td>
<td>49</td>
</tr>
<tr>
<td>Reliability</td>
<td>52</td>
</tr>
<tr>
<td>Topic Bias</td>
<td>53</td>
</tr>
<tr>
<td>Summary</td>
<td>55</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>56</td>
</tr>
<tr>
<td>WORKS CITED</td>
<td>59</td>
</tr>
<tr>
<td>WORKS CONSULTED</td>
<td>62</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>64</td>
</tr>
</tbody>
</table>
INTRODUCTION

Ever since the first composition student set stylus to clay tablet, we who teach composition have been blessed with the authority and cursed with the responsibility of evaluating student writing. Blessed because that grant of authority is an acknowledgment from our colleagues, our profession, our institutions, and our students that we can be trusted to fairly and accurately judge the quality of other people's writing. Cursed because that burden of responsibility demands an intensity of effort and an investment of time that are frequently intimidating and always exhausting.

Although there are a great many factors that contribute to the burden of evaluation, they can be broken down into two basic categories: (1) the goals of evaluation, and (2) the problems of evaluation. The burden is made heavier and the task of evaluation made more complicated by the fact that these goals and problems interact with each other both within and across categorical boundaries. However, for purposes of analysis, it can be said that the strategies we employ to solve the problems are the means by which we seek to serve the goals. More concretely put, when I count the grammatical/mechanical errors in a student's paper to ascertain whether or not that paper meets a correctness standard, I have used a particular strategy (counting) to address a specific
problem (the need to meet particular conventions) to achieve an evaluative goal (passing or failing the standard).

Teachers would not find the burdens of evaluation particularly onerous if we had but one goal to achieve and one problem to resolve in achieving that goal. However, as thorough evaluation has at least four goals—diagnosis of writing problems, directions for solving those problems, feedback on teaching skills, and assignment of grades—and presents a plethora of problems—among which are validity, reliability, and teacher biases toward topics—and as all of these elements interact with one another, evaluating student writing is obviously a far from simple task. Indeed, the task is so complicated with respect to the simple goal of giving grades that in *Grading Student Writing: A Plea for Change*, W. U. McDonald felt compelled to say that "[w]e are forced to reduce a complex set of observations and responses and assessments to a single symbol, the letter grade, a manifest impossibility for a composition course" (155).

Whether one views McDonald's statement as gospel or hyperbole, most composition teachers must, in the final analysis, assign grades to individual papers during the semester and to students at the end of the semester. This necessity has led teachers and researchers into a vast search for workable, usable evaluation strategies and systems. Some of these—such as the Ten Major Errors and Correctness Standard in place at Iowa State
University—lead to quantifiable data that can be used to determine students' writing success. Other systems—such as the Content, Material, Organization, and Expression grid used by many ISU composition teachers—reflect an effort to create discrete categories of writing features that can be qualitatively applied to any piece of student writing (i.e. Superior, Above Average, Average, Below Average, and Shoot Yourself). And still other evaluation systems—such as general impression ranking—reflect an attempt to view a piece of writing as a whole and evaluate it as a whole according to how well it meets a predetermined set of criteria designed for a specific writing assignment.

I have used versions of each of these strategies, and while each has its particular strengths, weaknesses, and appeal, I am most intrigued by those that fall into the third category, a category generally known as holistic evaluation. That attraction derives from what I regard as the two greatest strengths of such systems. First, holistic systems evaluate wholes rather than parts of wholes in the effort to judge and qualify student writing. Second, holistic systems generally have a collaborative element—i.e. more than one person reads and evaluates a paper.

I am, and have been, so intrigued with this approach, that, in cooperation with Walden Miller (like me, a graduate student in ISU's English department), I have been using a variation of it for
the past three semesters—a variation that Miller and I have dubbed Collaborative Holistic Evaluation (CHE).

Although CHE provides us with ways to solve most of the problems inherent in evaluating student papers, and those solutions enable us to achieve the goals of evaluation, a complete explication of the system is beyond the time and space limitations of this thesis. Therefore, I will focus on only one evaluation goal, three problems, and how our system works theoretically and practically to help us resolve these problems and achieve this goal.

More specifically, in Chapter One, I will discuss the goal of grading and the difficulties inherent in determining grades, and will also define and discuss the problems of validity, reliability, and topic selection as they inhere in evaluation in general and grading in particular. In Chapter Two, I will discuss the theoretical underpinnings of collaboration and holism, and describe the essential components of holistic evaluation systems. In Chapter Three, I will describe the system Miller and I have developed, and will discuss how it works to resolve the problems set out in the earlier chapters. Finally, I will offer some suggestions and directions for further research.
CHAPTER ONE: GRADES AND PROBLEMS

Before diving into the subjects of grades and problems, I must discuss the philosophical principles Miller and I share with regard to the nature of grades, and must briefly describe our theoretical approach to teaching writing. Additionally, I must explain the obligations our principles and theoretical approach impose on us with respect to assigning grades. Without this information, the reader of this thesis could easily get lost in the twists and turns of the explanations of the evaluation labyrinth.

To us, as to many, if not most, composition teachers, grades are highly subjective and relative, not objective and absolute. Our philosophical belief in this is predicated on our experience that different teachers will give different grades to the same paper, and that those teachers will each be "right" from their different perspectives. There is widespread disagreement about the definition of "good writing," and there is equally widespread disagreement about the key features of specific kinds of writing.

One outgrowth of these disagreements is that there are a number of theoretical approaches to what writing is and how best to teach it. Pedagogically speaking, Miller and I are both proponents of the process approach to teaching writing because we believe it fits our philosophical notion that writing is subjective in
nature, and the process approach emphasizes developing a variety of writing tactics to deal with the complexity of creating written discourse—i.e. this approach is cognizant of the subjectivity of creating writing, which matches nicely with the subjectivity of evaluation and grades.

We cannot, however, grade our students on their writing processes, for as Edward White says:

We make ourselves foolish if we ignore the fact that writing is a product as well as a process. Every student turning in a paper to be graded, every scholar producing a paper for delivery or publication or promotion knows perfectly well that writing is an important and measurable product. (Post-Structural 188)

Thus, we must grade products, and we must strive to make those grades relatively fair to ensure that students who have demonstrated equal competence in their writing get the same grade, and relatively accurate to ensure that grades reflect judgments of quality based on some definable, theoretically supportable set of criteria. To do less would be unfair to our students and to ourselves.
Grades

Over the past two decades, a great many college composition teachers have adopted the process approach to teaching freshman writing courses, and in so doing they have shifted the focus of daily classroom work from the analysis of completed texts written by professional writers (the traditional analytical approach) to the analysis of the evolution of texts as they are developed by both professional and student writers. Such process oriented classrooms concern themselves with the stages of writing— invention, drafting, revising, and editing—and the strategies, techniques, and tactics that can be used to accomplish these tasks. Among other hoped for benefits, adherents of the process approach seek to demystify writing by separating it into discrete kinds of problems (e.g. generating ideas, selecting details, choosing organizations) and providing strategies (e.g. free-writing, heuristics) to solve those problems.

Although I believe that the process approach is superior to the old analytical approach I grew up with, and although many of my students have commented that this approach has helped them to at last understand and improve their writing, the process approach poses a particular problem when the moment of assigning a single letter grade to a paper arrives. Simply put, how can I, or any other process teacher, focus classroom time on specific writing problems and strategies, then turn around and assign a grade that
reflects not only a student's skill in using those strategies to solve those problems, but also the student's skill in creating a whole paper?

For many of us, and for many of our students, composition courses would be far more pleasant if we could grade papers based on students' mastery of specific invention techniques or revision strategies or editorial conventions. However, as we know, and our students suspect, a piece of writing is not merely the sum of techniques applied, conventions used, and standard spelling and punctuation. The quality of a piece of writing also depends on its content, diction, tone, style, and organization. The quality of a piece of writing depends on its syntactical maturity, cohesive unity, pronoun usage, verb usage, and more. Indeed, texts are still so mysterious that after analyzing the relationships among holistic scores and twenty-four text variables, Cary Grobe was moved to say that "the ability to write . . . seems to be composed of complex sets of multivariate effects which act together in a presently unrecognized fashion" (85).

Stripped of jargon, Grobe is saying that the act of writing and the text produced are so complicated that we do not yet understand them. An automatic corollary of this is that writing teachers do not yet have a mathematically precise formula for evaluating student papers and assigning grades, for we do not know all of the variables involved, nor do we know how they influence
each other. We know only that there are many skills intrinsic to the act of writing and many features intrinsic to the quality of final texts. And somehow, we must judge all of these skills and features when we evaluate student papers and arrive at the grades we put on those papers.

The difficulty of assigning grades to papers caused by our lack of knowledge is exacerbated by a number of other problems too lengthy to list, let alone discuss, here. However, there are two definitional problems, mentioned earlier, so fundamental to evaluating student papers that they must be addressed. Those problems are:

- defining "good writing," and
- defining the key features of specific kinds of writing.

In *Composition Theory in the Eighties: Axiological Consensus and Paradigmatic Diversity*, Richard Fulkerson makes an argument that composition teachers in the 1980s came to "a significant consensus" about what constitutes good writing—i.e. what we value in writing and "what we want student writers to achieve as a result of effective teaching" (411). According to Fulkerson, we are now arguing about the means, how we can best teach our students to produce good writing, rather than debating the ends, what constitutes good writing.

The ends Fulkerson says we largely agree on are rhetorical ends that "value 'effectiveness,' audience awareness, persuasiveness,
and contextual flexibility . . . 'whatever works is right'" (409-10). To buttress his argument, he differentiates rhetorical ends from expressivist ends (authenticity, honesty, self-awareness), mimetic ends (accuracy in reporting), and formalist ends (correctness in style, structure, and format), and quotes several widely-known, highly respected composition authorities to show "the emerging rhetorical consensus on what constitutes good writing" (414). He goes on to show how textbooks, books on teaching writing, the increasing emphasis on audience, the awareness of writing as a social act, and programs featuring writing across the curriculum all reflect an axiological consensus that "good writing, the sort of writing that we hope to enable students to produce, is contextually adapted to, perhaps even controlled by, its audience (or discourse community), addressed or invoked, or both" (417).

Although he acknowledges that good writing always serves expressivist, mimetic, and formal ends, and "that a good deal of potential counter-evidence exists against my assertion of greater axiological unity today than a decade ago" (424), Fulkerson fails to be persuasive in his argument that we have reached a consensus in defining what good writing is because he confuses the concept of defining what something is with the concept of defining what something does; in Aristotelean terms, he offers only the final cause while arguing it is an essential definition. An examination of the authorities he quotes illustrates this quite nicely:
Maxine Hairston [says] . . . "[P]eople whose business it is to evaluate writing almost all agree that the key element of good writing is that it communicates effectively with the readers for whom it is intended . . . ."

Richard Marius says . . . "student writers need to learn that they have to interest someone."

Linda Flower tells students, ". . . a writer has to communicate that understanding so a reader will see what the writer meant." (414)

"Communicates effectively," "interest someone," and "has to communicate" are all phrases that express what "good writing" does, not phrases that say what "good writing" is.

Superficially, it may seem that I am splitting rhetorical hairs in differentiating between definition and description. However, it is a significant distinction, especially when the time arrives to justify the grade I give a student paper. I cannot simply write or say to a student that a paper "doesn't communicate effectively"; I must state why it fails to do so, and to do that, I must tie the comment to specific features of the text. A memo, a proposal, a research paper, a personal narrative, and a critical essay may each be "good writing," but the writing in each of them will be so different from the writing in the others that no set of features will be common to
all of them—i.e., there will be no shared definition of good writing, only shared descriptions of the impact of "good writing."

Thus, in teaching our students and grading their papers, we have no workable definition of what "good writing" is, no universally applicable set of features, standards, or guidelines by which to judge all student papers. We have only task-specific criteria that rely on what a paper's purpose is and how well that purpose is met as judged by its impact, and we look for features that we deem appropriate to accomplishing that purpose.

At first blush, it would seem that knowing the important features of specific kinds of writing, or even using the general feature analysis set out earlier, would provide composition teachers with the analytical tools that would enable them to fairly judge and grade student papers. Indeed, it would seem that if we accept Fulkerson's argument that rhetorical aims are the philosophical underpinning of "good writing," and that "good writing" is whatever works, then we have a general standard to apply to those features, and determining grades is a simple matter of applying the standard to the specific or general features. Unfortunately, such feature analysis is subject to the same sort of definitional problems that make defining "good writing" impossible.

Sarah Freedman provides a fine example of the problem of defining the features of good writing in her article Why Do Teachers Give the Grades They Do? In that piece, she reports the
findings of a study she conducted in an effort to answer three questions: (1) the question in her title, "(2) are there any specific, definable parts of student papers that influence teachers? (3) and if there are, which of the parts influence teachers most?" (161).

To conduct her study and answer her questions, Freedman "rewrote student papers to be weak or strong in four broad, but pedagogically interesting, areas: content, organization, sentence structure, and mechanics" (161). Before these areas, or features, could be used to judge the papers, she had to define each one "in a way that would make it discrete from, or independent of, every other area" (161). It is in her effort to define one of these areas or features—content—that the problem of feature definition is made evident.

Briefly, content was the development of, and logical consistency between, the ideas. It had nothing to do with the absolute quality of the ideas. . . . So when I say good teachers valued content most of all, I mean that they valued the development and logical presentation of the ideas, not necessarily the ideas themselves. (161)

To me, "development" and "logical presentation" are as much matters of expression and organization as they are of content. In fact, Freedman herself makes "order" and "transitioning" elements
of organization, yet she does not define the difference between "order" and "logical presentation," nor between "transitioning" and "development." In short, her definitions of key features are not discrete, and are certainly not universal. They are her definitions.

My intent here is not to denigrate the work done by Freedman, Fulkerson, and the many other scholars, researchers, and teachers who have sought to introduce some sense of order into composition studies and evaluation. Rather, they are to be commended for their efforts to undo the Gordian knot of writing and writing evaluation, their attempts to find philosophical consensus, and their efforts to create analytical rigor in a field that resists analytical unification. Their work is important and significant, if for no other reasons than it defines and describes the serious problems we who teach and grade writing face, and it reveals how complex the tasks of writing and evaluating are.

Nevertheless, despite the work of hundreds of composition teachers and researchers over the past thirty years, "[c]ollege writing research . . . has not told us much about exactly what it is that teachers value in student writing" (Faigley 359). Our "definitions of good writing are either circular or absent altogether" (Faigley 359-60), and our definitions of the key features of writing are hopelessly muddled and subjective. When our lack of standardized definitions is combined with our ignorance of all of the variables involved in writing and grading, then it would seem
that McDonald may speak the truth—assigning a letter grade for a paper or a course is an impossible task. And, in one of the world's little ironies, it is an impossible task that most of us must perform dozens or hundreds of times a semester.

Given that assigning letter grades is at least highly problematical, we can still seek to give grades that are relatively fair and accurate. However, any system or strategy for achieving even this limited goal must deal with the problems of validity, reliability, and topic bias, matters to which I now turn.

Problems

Validity

In its simplest meaning, validity is the notion that a system of measurement or judgment should actually measure or judge what it purports to measure or judge. "A valid measurement assesses what it claims to assess. So, a valid writing assessment would be sensitive to a writer's 'true' abilities" (Charney 65). Analogically speaking, a valid system of writing evaluation in the hands of a competent evaluator would measure the quality of a piece of writing as accurately as spectrographic analysis conducted by a competent gemologist would measure the quality of a diamond, and the grades given by the former would be as trustworthy as the grades given by the latter.
However, as with everything about writing evaluation, the concept of validity is more complicated than its simple definition indicates, for there are four kinds of validity involved in the overall notion of validity: predictive validity, content validity, construct validity, and concurrent validity. Additionally, there is the necessity of determining who is competent to accurately apply a fully valid evaluation system once that system has been devised.

Brian Huot neatly summarizes the four types of validity important to writing evaluation systems in *Reliability, Validity, and Holistic Scoring: What We Know and What We Need to Know*:

- **Predictive validity** has importance for placement testing, where ratings given should be able to predict the amount of success a student will have in a specific English course.
- **Concurrent validity** is the ability to correlate one type of testing with another (cite omitted).
- . . . . Content validity assumes that the assessment instrument contains the necessary procedures to truly measure for its intended purpose . . . [and does] allow for the real measurement of writing ability. **Construct validity** insures the theoretical soundness of an assessment procedure. "The construct validity of a test is the extent to which the test may be
said to measure a theoretical construct or trait" . . . a student who scores well
. . . would be considered a competent writer,
and a student who scores poorly would be
defined as an incompetent writer. (206)

It must be noted here that Huot is addressing the validity
requirements of evaluation systems used in the large-scale testing
environment of placement exams, entrance exams, basic skills
exams, and exit exams (208), and that his article is a critique of
holistic evaluation in that environment. Furthermore, the source
for his definitions of the four kinds of validity and their purposes is
the Standards for Educational and Psychological Tests put out by
the American Psychological Association in 1974. While the nature
and age of his source might call into question the validity of Huot's
adoption of these definitions and purposes, and his use of them to
critique any system of writing evaluation, I find his explication
useful because three of the four kinds of validity he sets out—
predictive, content, and construct—have corollaries in the small-
scale environment of the classroom. Also, those three kinds of
validity illuminate grading problems that any system of evaluation
designed for classroom use must cope with.

The classroom corollary of predictive validity in the large-
scale test environment is not, as one might think, that a student
who does well on one assignment is likely to do well on the next.
Instead, the corollary is that a student who passes a first semester writing course with an 'A', 'B', or 'C' grade is likely to pass a second semester writing course even if—or especially if—the teacher for the latter is not the same as the teacher for the former. If a student receives 'D' in the course, then he or she is unlikely to pass the second semester course, and if the student receives an 'F', then he or she will not pass the second semester course because he or she will not be permitted to take that course. Additionally, grades of 'A', 'B', or 'C' in both courses (or one course if only one is required) represent a prediction that a student will be able to write successfully for courses in other academic areas. The problem of predictive validity in the classroom, then, is how to ensure that the evaluation system used will appropriately distinguish between those students who should be passed on and those who should be held back (or at least forewarned).

Content validity in the writing classroom is largely the same as content validity in the large-scale testing environment, and the procedures used to evaluate classroom writing must be designed to accurately evaluate writing ability as it is demonstrated in particular papers so that the grades assigned to those papers reflect the quality of demonstrated writing ability. In other words, the procedures must eliminate (or at least adjust for) the influences of variables extraneous to the actual writing (e.g. the health of the evaluator, the physical environment, evaluator bias toward a student
or topic, etc.), so that the evaluator is focused on the writing and the grade given is predicated on how well that piece of writing meets the applicable criteria. Additionally, the evaluation procedures must in some way ensure that evaluators are in fact using the applicable criteria rather than some set of inapplicable, irrelevant, or unimportant criteria.

Unlike predictive and content validity, the classroom corollary of large-scale construct validity is not quite so evident, especially as large-scale construct validity depends on repeated use of the same testing instrument to establish its surety (Huot 206). Nevertheless, there is a corollary if the concept "theoretical construct or trait" is taken to be equivalent to the concept of "key feature" used earlier in the discussion of grades. If those concepts are equivalent, and I think they are, and if it is possible to describe the key features of specific kinds of writing, which I think it is, then a classroom assignment (the corollary of the testing instrument) can have construct validity. Moreover, if it is possible to recognize and qualify the success of a student writer at demonstrating those features, which I think can be done, then the procedures used to assess student writing can have construct validity as long as those procedures accomplish two aims: (1) they must be designed to ensure that evaluators know and look for the instantiation of the key features, and (2) the same assignment and evaluation procedures must be used repeatedly. Thus, the
problems of construct validity in classroom evaluation and grading
that must be resolved are, first, that the evaluation system must
focus on described key elements, and, second, that the same
assignment and evaluation system must be used repeatedly.

Concurrent validity in the large-scale environment—"the
ability to correlate one type of testing with another"—has no
corollary in the classroom and no corollary in evaluating papers
from the classroom. In fact, such validity could pose a danger to
fair and accurate grading because seeking this kind of validity poses
the danger of biasing teachers towards students. Put another way, a
student who gets a 'C-' on the first paper of the semester should
not have to overcome that grade to get an 'A' on the second paper.
The two papers should be judged and graded independently so as
to maintain content and construct validity. In short, concurrent
validity is anathema to accuracy and fairness within the classroom
grading environment.

As mentioned earlier, the validity of an evaluation system
depends as much on who applies it as on the system's design. In
the classroom setting, this necessity generates a fundamental
problem because "instructors assigned to the task [of teaching
composition] find it impossible to agree on what constitutes good
writing, how it should be taught, or even if it should be taught"
(Cohen 356). Assuming, for the moment, that agreement can be
reached that composition should be taught, and that approximately
similar teaching approaches can be used, and that the definitional problem can be solved with respect to specific assignments by establishing sets of key features, we are still confronted with the problem of ensuring that evaluators are competent at identifying those features and assessing a student's success at meeting them (McColly, Cooper, Charney, Freedman, et alia). We must, in essence, have some means of validating the expertise of evaluators, both in general (i.e. their background and experience) and in regard to a specific assignment, for an evaluation system to be validly applied.

Thus, the simple notion of "valid evaluation" and its manifestation in fair and accurate grades poses a variety of problems for writing teachers who must assign grades to student papers. An evaluation system must be such that teachers take into account the predictive validity of course grades, and the content and construct validity of individual paper grades, while avoiding the subtle bias inherent in concurrent validity. Additionally, the system must have methods for establishing the expertise of teachers, their validity, to act as evaluators, judges, and graders.

The complexity of arriving at fair and accurate grades is not solely a matter of the validity of the grader and the evaluation system used. That system and those using it must also be reliable, the subject of the next section.
Reliability

In his 1984 article, *Holisticism*, Edward White sets out problems in reliability as he defines that concept as it pertains to the large-scale testing environment.

Reliability in scoring is a complex subject, with many shapes and forms. No test, of course, is wholly reliable, since student performance will change from test to test for reasons that have little or nothing to do with the test. Scoring reliability is only one part of this complicated problem, but it is the part most directly controlled by those in charge of the test. Since reliability is in a sense a technical term to describe fairness, or simple consistency, good testing practice aims for the highest reliability that can be reached. . . . unfair and inconsistent scores are meaningless, and meaningless scores . . . are not worth anything at all. (403)

As with validity, reliability in the large-scale test environment has its corollaries for evaluating and grading student papers from the classroom environment. A student's writing quality will vary from assignment to assignment because of outside influences over which writing teachers (and frequently the students themselves)
have no control, and so there is a student component to the reliability of any given grade. Simply put, a student may not put forth the best effort on an assignment, and the writing produced for that assignment will be less than an accurate reflection of the student's true ability. However, this student component of reliability, important as it may be to classroom teaching, is not significant in determining the reliability of an evaluation and grading system. Indeed, a reliable system would give a paper the grade it merits irrespective of this student component. What is important to that system's reliability is that it be consistent. As Davida Charney says, "[a] reliable measurement is capable of replication under equivalent conditions. So, a reliable method of assessing writing ability would yield a consistent judgment of a student's abilities if applied again, all else being equal" (65). In other words, a system of evaluation is reliable, and a grade is reliable, if two or more readings of the same paper, whether done by one reader or multiple readers, result in the same grade. Thus, the problem of reliability in an evaluation system is simply that the system must encourage consistency across readings.

Since both the reliability and the validity of a grade are concerned with the accuracy or fairness of that grade, it is easy to confuse the two, a confusion which Huot traces to a misunderstanding of the role of the criteria used to evaluate student writing: "However, any given set of criteria, no matter how
effective at providing rater agreement [reliability], cannot by itself insure a true assessment of writing quality (validity)” (205). What Huot is driving at is that simple agreement as to the quality and grade of a piece of writing—whether the result of multiple readings by one grader or single readings by multiple graders—is not sufficient to establish the accuracy or fairness of that grade. In other words, that I give a paper a 'B' after each of two separate readings, or that both I and someone else each give a paper a 'B' after one or more readings, establishes, at best, that I or we are reliably applying the same set of criteria in all readings. The 'B' and the agreement do not establish that the criteria themselves are valid, nor that the grade arrived at reflects fairly the quality of the writing. In a sense, the validity problem of criteria is that they must "truly" assess writing quality so that different papers of equal quality written by different students of equal ability receive the same grade. The reliability problem of criteria is that they must be applied consistently so that the same paper receives the same grade after different readings. For teachers who have to give grades, validity problems are problems of abstraction, while reliability problems are problems of application, and nowhere are the application problems more evident than in teacher response to student topic selection, the last of the three problems I will deal with.
Topic Bias

A common feature of many composition courses is that at some point in the semester, students are required to choose a topic to write about. The assignment may be to write a narrative account of a significant personal experience, to write a report based on research, or to argue a position on an important issue. Whatever the nature of the assignment and the skills necessary to accomplishing the assignment, however, the student must choose the general subject and frame the specific topic about which he or she will write. Indeed, learning to choose subjects and to focus in on topics are themselves important writing skills, whose demonstration will be judged in the evaluation and reflected in the letter grade.

With that student freedom to choose, and the obligations it imposes on the student, come the teacher's obligations to accept student choices and positions, and to avoid imposing personal biases and opinions on student papers. While this may, at first, seem to be an obvious problem, and one readily solved by admonishing ourselves to "look at the quality of the paper, not the nature of the subject or position," it is neither always obvious, nor easily solved.

Lad Tobin addresses the problem of teacher bias toward topics (and students) in *Reading Students, Reading Ourselves*: 
Revising the Teacher's Role in the Writing Class. In this article, Tobin describes how he organized and focused a favorite student's paper when discussing it with colleagues. Upon rereading the paper after that discussion, and discovering his "appropriation" of the student's text, he also discovered that the student's position in the paper was akin to his own position on the same subject.

I began to see how much her whole argument echoed my own ideas... ideas to which I have a strong ideological and personal commitment. ... So it only makes sense that I would be pleased and excited to see that my student's writing supported and even validated my own positions and, therefore, that I would make her argument more eloquent and sophisticated than it really was. (334)

Although Tobin does not say what grade he gave the paper, I think it reasonable to assume that it was a high grade, especially as he describes the paper as "a particularly insightful essay one of my students wrote about the relationship between thought and language" (333).

Irrespective of that grade, and far more important here, is that Tobin straightforwardly admits that until his rereading of the essay—presumably done after it had been graded—he had not been consciously aware that his pleasure and excitement at the first
reading had at least in part been caused by his agreement with the student's position. In effect, his evaluation of the paper reflected his favorable predisposition toward the topic and the student's position on that topic, not just the quality of the writing. In fact, he implies that his bias may have been more important than the quality of the writing when he says, "in general the writing seemed much flatter and more prosaic that [sic] I had remembered it" (334).

Unconscious biases are not, however, the only problematical form taken by teachers' predispositions towards topics and positions. There is also the problem generated by the pretext that conscious biases can be laid aside. Tobin illustrates this problem, too.

We are not unaware that we bring to our teaching of writing and our reading of student essays strong beliefs and biases. We know, for example, how we feel about abortion and gun control, how we respond more favorably to some rhetorical strategies than to others, even how we like some students much better than others. But we conveniently forget those issues and pretend we can willingly suspend those beliefs and disbeliefs. We see ourselves as neutral, objective, open-minded. We give
each student an equal chance. We are ready to like essays on any topic in any mode. . . .

This paradigm of the teacher-as-objective-reader fails to do justice to the complexity of the reading and writing processes . . . (336)

Implicit in Tobin's discussion is the truism that we cannot suddenly divest ourselves of the opinions, predispositions, and biases we have when we don the role of evaluators and graders. To this can be added an emphatic, "Nor should we." Much effective writing depends, in part, on an appeal to the emotions, and responses that arise from biases are inherently emotional. Students deserve to be rewarded for using our buttons to increase their effectiveness as writers, which means that we must allow those responses not merely to exist, but also to play a part in our evaluations and grades.

Nevertheless, we cannot allow our biases and their emotional impacts to outweigh our evaluation of a student's writing as writing; we cannot allow, as Tobin did, our favorable or unfavorable personal responses, our agreements or disagreements, to supersede our judgment of the quality of a student's text; we cannot, in short, let what is in us dominate our assessment of what is on the page. To do so would render the grade inaccurate and unfair, and make our grading invalid and unreliable. Thus, a system of evaluation must provide some means for ensuring that teacher biases toward a topic
or position, and the emotional responses deriving from those biases, do not determine assessments of quality and grades.

Summary

As stated earlier, grades are not the only goal of evaluation, nor are validity, reliability, and topic bias the only problems that an evaluation system must find resolutions to. However, in an increasingly number-happy, bottom-line society, grades are perceived to be important, both inside and outside the ivory towers, and therefore are important. Bluntly put, whether letter grading is possible and desirable or impossible and undesirable is irrelevant—with rare exceptions, it must be done. And, since the goal must be achieved, it is up to teachers not only to devise ways and means to achieve it, but also to do so with some sense of fairness and accuracy. These requirements, in turn, demand that our grades reflect a true assessment of demonstrated writing ability (validity), be consistently fair (reliable), and that we avoid overvaluing or undervaluing emotional responses that exist within us (topic bias). Moreover, to be acceptable to members of the composition profession, other educators, and administrators, any system of evaluation used to arrive at grades must be predicated on sound theoretical principles and pragmatic applications that solve the problems of validity, reliability, and topic bias.
Collaborative Holistic Evaluation is one such system. However, before I describe CHE and demonstrate how it functions, I must explicate the theoretical underpinnings of collaboration and holisticism, and discuss how they resolve the problems set out.
CHAPTER TWO: THEORETICAL UNDERPINNINGS

As I indicated in the INTRODUCTION, collaboration is usually a component of holistic evaluation systems. In fact, it is a universal component of all such systems designed for use in large-scale testing situations (White, Holisticism 405). Thus, it may seem unnecessarily reductionist, if not redundant, to deal with collaboration theory as a separate entity. However, because CHE is a system designed for the small-scale classroom situation, and is but one component of a much larger, more encompassing teaching system, I feel it is important to address collaboration separately, especially since I will be referring to other parts of that larger collaborative teaching system later in this thesis.

Having explained why I am separating collaboration and holismicism, I will now turn to a brief discussion of the theoretical foundations of collaboration. From there, I will go to a similar discussion on holismicism, and thence into a brief explanation of the elements of holistic evaluation systems.

Collaboration

Within the confines of the writing classroom, formal collaboration is usually thought of and used as a student-to-student pedagogical tool—a means by which students work together to
develop and improve their writing skills and their final products. In the past fifteen years or so, it has become so popular that, as Harvey Kail notes in *Collaborative Learning in Context: The Problem with Peer Tutoring*, "students are being required to work on their writing together, commanded to learn from each other; they must collaborate. We have begun to insist on it" (594).

Oddly, however, formal classroom collaboration among composition teachers has not kept pace with its use among students, despite teachers' "long tradition of [informal] collaborative learning outside the classroom" (594), and despite the growing frequency of collaboratively authored journal articles directed at our professional peers. The reasons for this, interesting though they might be, is not important here. What is important are the notions that composition teachers have as much to gain from collaboration as composition students, and that the theoretical bases of collaboration are the same for us as teachers as they are for our students as students. For both groups, the definitions and interworkings of knowledge and the community of experts provide the theoretical support for collaboration.

In his seminal work *The Structure of Scientific Revolutions*, Thomas Kuhn argues that scientific knowledge is not truth—or even a close approximation of truth—as logical empiricists would have it, but a paradigm of "the entire constellation of beliefs, values, techniques, and so on shared by the members of a given
community" (175). Knowledge, then, is a social artifact whose validity rests not upon any absolute foundation, but upon the shared conviction of a group of experts in a field—a community. Simply put, knowledge is knowledge because they—the community of experts—say it is knowledge. Moreover, that community of experts is a community because the members share a particular knowledge paradigm, "they have undergone similar educations as professional initiations; in the process they have absorbed the same technical literature and drawn many of the same lessons from it" (177). Finally, Kuhn argues, it is this community that not only produces and validates knowledge, but also trains the successors to that knowledge (177-78).

Put another way, Kuhn is saying that knowledge is a matter of agreement among experts, a structure of shared understanding made possible through acts of collaboration among experts. Quite naturally, such acts of collaboration require communication, and such communication is always some form of conversation. As Kenneth Bruffee says in Collaborative Learning and the "Conversation of Mankind,"

We establish knowledge or justify belief collaboratively by challenging each others' biases and presuppositions; by negotiating collectively toward new paradigms of perception, thought, feeling, and expression;
and by joining larger, more experienced communities of knowledgeable peers through assenting to those communities’ interests, values, language, and paradigms of perception and thought. (646)

If Bruffee is correct, and “challenging,” “negotiating,” and “joining” are the forms that conversation must take to establish knowledge, then he is equally correct in saying that “[c]ollaborative learning provides a social context in which students can experience and practice the kinds of conversation valued by college teachers” (642). By analogy, then, collaboration among teachers provides the social context in which we can converse and establish exactly what kinds of writing we actually do value as members of the community of composition experts. Additionally, the more exactly we know the nature of a writing assignment, the more we have discussed it with other composition experts and arrived at a shared understanding of what writing features we are looking for in that assignment, the more able we are to establish the proper criteria for evaluation and explain those features and criteria to our students. More importantly, the better our understanding of the features and criteria, the more accurately and fairly we can evaluate and grade student papers according to the criteria we have developed and espoused.
Thus, for writing teachers, formal collaboration provides a means by which we can develop our knowledge of specific writing assignments and increase our abilities to accurately apply the criteria we have developed. Not surprisingly, these improvements can lead to greater reliability and validity of the grades we assign to individual papers, and to a significant reduction—if not elimination—of the influence of topic bias on those grades. However, before addressing how these effects can result from collaboration, it will be necessary to place collaboration within the framework of holistic evaluation, and to discuss the theoretical underpinnings of holisticism.

Holisticism

The theoretical foundation of holistic evaluation has two basic roots—one in the manner of how student papers are viewed, the other in the collaborative nature of holistic systems. I will briefly explicate these two roots, then provide a brief description of the elements of holistic evaluation systems.

In Chapter One, I argued that composition teachers do not know all of the skills involved in writing, and neither do they know all of the features that make up a written text, and that because of these we do not have a mathematically precise formula for evaluating student papers. In essence, that argument is a rejection of the analytical approach to evaluating student writing—a rejection
predicated on the premise that analytical reductionism (i.e. analysis of the parts to understand the whole) is workable if and only if all of the parts can be identified and their interactions known. As that sort of complete knowledge does not exist with regard to written texts, any purely analytical system of evaluating and grading student papers will automatically be inadequate to its task, for such a system cannot "add up" an incomplete column of "parts" and arrive at a "total" that will equal the whole of a text. At best, any such system can arrive at only a part of the total, only a partial score of the whole. And that partial score will not, indeed cannot, reflect the impact of a text as a whole.

Obviously, however, composition must have some method for evaluating student-produced texts as wholes, and some method for arriving at whole grades, and this is the goal of holistic evaluation. As Edward White says,

Thus, holism, with its emphasis on evaluation and response to student writing as a unit without sub-scores or separable aspects, presents itself in opposition to multiple-choice testing on the one hand and analytic approaches to writing on the other. It is the most obvious example in the field of English of the attempt to evoke and evaluate wholes rather than parts. (Holisticism 400)
Implicit in White’s statement is one of the theoretical foundations of holistic evaluation. Quite simply, that foundation is that the whole of a text is greater than the sum of its parts, especially as not all of the parts and their interactions are known and understood. In other words, a student text is singular in nature, and as such must be responded to and evaluated as a unit, “not merely as a collection of scraps and parts” (409). As Charney neatly summarizes it, “the common assumption behind all [holistic evaluation systems] is that a valid assessment of writing ability includes a natural human response to a writing sample” (69).

The other theoretical foundation of holistic evaluation is reflected in the collaborative nature of its methodology, a feature of such systems that will be detailed a bit later. For now, it is sufficient to note that collaboration is a key element of such systems, and to explicate how collaboration functions theoretically in holistic systems.

Earlier in this chapter, I briefly discussed Kuhn’s argument that all scientific knowledge is the result of collaboration and Bruffee’s argument that knowledge in the writing community is the result of challenging, negotiating, and joining the community of experts through conversation. If these arguments are valid, and knowledge is the product of conversation among experts, then it is a short step to the further conclusion that evaluating students’ academic products is largely a matter of comparing those products
to the body of knowledge established by the collaborative efforts of the relevant community of experts and judging the quality of those products by standards created by that community. Put more simply, a student's efforts should be judged not by the idiosyncratic standards of the individual, but by the communal standards of the community of experts. This, in turn, means that standards of evaluation, if they are to be adequate, must be generated by and accepted by the community of experts or some subset of that community that is larger than the individual. As Kuhn says, "the solutions that satisfy [the scientist] may not be merely personal but must instead be accepted as solutions by the many" (168).

Thus, the theoretical roots of holistic evaluation are that (1) a holistic evaluator reads and responds to a student paper as a whole, and (2) that the response and quality judgment are predicated on standards collaboratively generated by the evaluator's community of experts.

Methodologically, these theoretical foundations are manifested in seven key features common to all systems of large-scale holistic evaluation. As set forth by White (Holisticism 404-05) and Charney (67), those features are:

1. that essays be read at the same time and place by all readers,

2. that a criteria guide or rubric of important writing features be developed specifically for the essays,
3. that readers be familiarized with the rubric and calibrated using sample papers until they reach close agreement on the scores given to the sample papers,
4. that checks on consistency of scoring be performed by head readers during the evaluation session,
5. that two readers score papers independently—i.e. without knowing the scores given to papers by other readers,
6. that records of readers' scoring are kept with close attention to readers' consistency with each other, and
7. that essays be read quickly and scored impressionistically rather than analytically or diagnostically.

These features demonstrate that holistic systems in the large-scale test environment are designed to implement the theoretical foundations on which holistic evaluation rests. Features one and seven show the theoretical underpinning of responding to student papers as wholes, while features two through six reveal the collaborative base of holistic systems.

Summary

As shown in this chapter, collaboration is essential to generating knowledge, and holistic evaluation—with one root in collaboration and another in anti-reductionism—is a means of
judging texts as wholes rather than as parts of wholes. In the next chapter, I will describe the classroom system Miller and I have developed, show how it manifests the seven key elements common to holistic systems, and explain how our system solves the problems of validity, reliability, and topic bias set out in Chapter One.
CHAPTER THREE: COLLABORATIVE HOLISTIC EVALUATION

The system of evaluation that Walden Miller and I have devised came about as the result of two events that occurred toward the end of our first semester as graduate teaching assistants. The first was that we collaboratively designed and implemented a collaborative writing assignment for our students. The second event was that we collaboratively produced a syllabus and syllabus rationale for an assignment given to us in our graduate proseminar on pedagogical theory and practice. Thus, our system evolved from both sides of the teacher's desk, and as such was heavily influenced by our efforts to see collaboration from two perspectives—as students and as teachers. In a sense, it would be fair to say that our entire collaborative teaching system, including CHE, developed from mutually induced, mutually reinforced schizophrenia.

Although there are a great many elements involved in our teaching system, discussion here will center on only the two most important to this thesis: assignment design and evaluation methodology. The former is the classroom equivalent of the testing instrument of large-scale test environments, and the latter is the classroom corollary of holistic evaluation. As, chronologically speaking, assignment design precedes evaluation, I will discuss these in that order, and then address how CHE resolves the
evaluation and grading problems of validity, reliability, and topic bias.

Assignment Design

Recently, a former student brought me a copy of a writing assignment she had received from her second-semester composition teacher and plaintively asked me, "What does this (expletive deleted) teacher want?" After we discussed the assignment, and the student agreed to talk with her current writing teacher, the student commented that she now appreciated the clarity of the assignments I had given her and "wished all teachers were so understandable."

After the "warm fuzzies" induced by this student's compliments had passed, I reflected on the difference between my writing assignments and the one she had shown me from another teacher, and came to the conclusion that mine were so much "clearer" because they weren't all mine; they were the result of intense collaboration between Miller and me, and between me and my students. They were better assignments because they were the products of collaboration between professional peers—representatives of the community of composition experts—and collaboration between a teacher and his students. In essence, they were better because they were the results of "the sorts of conversation members of the community value" (Bruffee 640).
Although much has been written about "the best way" to design assignments and the features of "good writing assignments," the effects of assignment design on student writing are still not clearly understood. Indeed, as Charles Kinzer points out in Effects of Topic and Response Variables on Holistic Scores, even the effects of so apparently simple a matter as topic selection are not well understood because "there is as yet not a 'critical knowledge base' to clearly answer the question about the effect of topic and the demands placed on the writer by particular topics" (106).

In an effort to determine the effects of topic selection on the quality of student writing, Kinzer studied the holistic scores given to essays written on two topics by comparable groups of students. Among other things, he found that essays written on the topic that had more explicit demands were rated by evaluators as being better essays than those written on the topic that made more implicit demands (117). In short, students wrote better essays when the assignment was clear in its demands. Kinzer goes on to conclude that "a discussion and explanation of expectations, related to scoring, can alert writers to the necessity of addressing topic task demands that may not be explicitly stated" (118). In effect, Kinzer is arguing that teachers should discuss assignments with students to make explicit what may be only implicit in the text of an assignment, to read aloud from between the lines.
What Kinzer proposes is exactly what Miller and I do in two stages. First, we discuss a tentative assignment and arrive at agreement as to the topic, length, due dates, and especially the criteria for evaluation. One of us then drafts the text of the assignment, and we review and revise it until we are satisfied that it is as clear and understandable to us as we can make it; we create our knowledge of the assignment. In the second stage, we distribute the assignment to our students and discuss it with them until they indicate that they understand the assignment and the criteria by which their papers will be judged; we create the knowledge shared by the classroom community. Thus, our assignments—our testing instruments—are the products of a two-stage collaborative process that involves every member of the classroom community in generating the knowledge of what is expected in papers and how those papers will be judged. Students, then, know before setting pen to paper exactly what they must accomplish and the standards by which they will be judged.

Evaluation Methodology

After students have turned in their papers, Miller and I set a time and place to meet to evaluate papers. Although our usual place is The Commons in Iowa State's Memorial Union, we have met at a variety of other locales, all of which share three features:
• They are public areas with a background of "white noise" that serves to isolate us and insulate us from distractions.
• They serve coffee at a reasonable price.
• They have well-lighted tables large enough to hold stacks of papers, coffee cups, and ashtrays.

These three features enable us to satisfy the first element of holistic systems—that essays be read at the same time in the same environment by all readers.

Once we have coffee and have chatted about matters other than the papers, Miller and I review the text of the assignment and any notes made on the assignment sheet during class discussions, with particular attention to the criteria for evaluation. From this, we develop the rubric of key features specific to the assignment and discuss these features until we are comfortable with them—i.e. have reached agreement about what they are. This satisfies element two (development of specific criteria guide) and part of element three (familiarization of readers with the rubric) of holistic evaluation systems.

Up to this point, CHE exactly matches large-scale holistic systems. However, from this point of the process on, there are some significant differences between CHE and the systems described in Chapter Two, differences necessitated by the nature of the environment (i.e. classroom vs. large-scale) and by the goal (i.e. ranking vs. grading).
The first difference is that we do not use sample papers to calibrate ourselves at the beginning of the session. On the one hand, if we were to use actual papers to do so, then the selected papers would receive more attention and scrutiny than other papers, which could lead to an unfairly exacting examination of the selected papers. On the other hand, if we were to create sample papers, we would run the danger of creating models, which could subconsciously replace the rubric as the standard of evaluation. Thus, instead of pre-session calibration, we rely on post-session calibration, which I will detail later on.

With the rubric firmly in place, we quickly read each paper and assign it a grade on separately kept sheets where we also jot down quick comments about our impressions of each paper, impressions related to the established criteria. This part of CHE satisfies elements five (independent scoring) and seven (quick reading and impressionistic scoring) of holistic systems, and part of element six (record keeping).

Not satisfied are element four (consistency during the session) and part of element six (records kept with attention to reader consistency). Like the missing part of element three (calibration), these elements are satisfied in the post-session discussion, to which I now turn.

After we have independently read and graded each paper, Miller and I compare our grades and comments, and it is at this
point in the process that the calibration and consistency elements of holistic systems are satisfied. Papers that have equal grades or grades that are within one step of each other (e.g. a C+ and a C, or a B- and a C+) are passed over quickly because differences from zero to eight percent indicate that we are consistently applying the established criteria. Obviously, the larger the group of papers is that falls into this category, the greater our consistency. This component of CHE satisfies the consistency requirements of elements four and six of holistic evaluation systems.

Always, however, there are papers that have grades two or more steps apart, and these are the discrepant papers that serve to calibrate our grading. Each such paper is pulled out and re-examined. We compare our notes and grades to determine why we gave the same paper such different grades, and in most cases to date, we have found that one of us applied some set of criteria to the paper other than the set agreed upon at the beginning of the session. In such cases, it is always readily apparent which of us has erred, and usually apparent as to what the cause of the error is—emotional response, an element of holistic evaluation Donald Daiker addresses in The Student Essay as Dubloon: Discrepancies in Holistic Evaluation.

For his study, Daiker examined both the student paper that had received the greatest range of scores from a group of sixty-one holistic evaluators and the comments that those evaluators had
made about that paper. What he found was that "[t]he one issue which most consistently created disagreement was the paper's emotional impact" (134). Those readers who were moved by the paper gave it a high score, those unmoved gave it a low score. After cautioning the reader that emotional responses must be watched for, and arguing that such discrepancies show the need for appeals procedures by which students can challenge grades, Daiker goes on to say that discrepant papers serve to remind teachers "that none of us responds to any text in exactly the same way and that 'calibration' is not always the highest goal in human affairs" (140). While I would not disagree with Daiker about any of these conclusions, I feel it necessary to discuss a subtle flaw in his work that does not exist in our CHE system.

In Daiker's examination, the evaluators were not asked to provide, and did not provide, the personal background information that could have helped to explain their emotional responses. Rather, only responses to the text qua text were asked for and received. Obviously, an evaluator could have responded strongly from personal experience and allowed that experiential response to dominate his or her score. Such domination, as I say in the first chapter, cannot be allowed to supersede our evaluation of a student's writing as writing. Some means must be in place to eliminate this.
Fortunately, CHE has a built-in method to control for grades that are the result of internal emotional responses to the evaluator's personal experience. That control is to ask each other if there is a personal response, one more to what is inside the reader than to what is in the text. If, as occasionally happens, there is such a response, then it can be discussed and included as part of the evaluation. It will not dominate the evaluation, nor will it dominate the grade given to the paper.

As a result of such discussions and agreements, we are able to satisfy the calibration requirement of the third element of holistic systems, and thus incorporate all of the elements of holistic evaluation in our CHE system. Pragmatically speaking, this means that Miller and I have a workable methodology for grading student papers. Theoretically speaking, it means that we have a solid theoretical base for solving the problems of validity, reliability, and topic bias, matters to which I at long last turn.

**CHE and Three Problems in Grading**

**Validity**

As explained in Chapter One, there are three types of validity an evaluation system must demonstrate if the grades derived from that system are to accurately reflect a true assessment of true writing ability: predictive validity, content validity, and construct validity.
The predictive validity of a classroom evaluation system can best be demonstrated by the success or failure of students in later composition courses. Simply put, students who receive grades of C or higher in first-semester composition courses should be able to achieve grades of C or higher in second-semester composition courses taught by different teachers. Although I do not have any statistical data to show that C and above students from our courses uniformly or largely achieve that success, I do have anecdotal evidence from seven of my former students that indicates that CHE has predictive validity. Of those seven students, two received the same grade in their second-semester composition courses that they received in my first-semester course, and five got higher grades in their second-semester courses. To date, no student has reported receiving a lower grade. While this is far from the weight of evidence needed to prove the predictive validity of CHE, it nevertheless lends support to CHE’s predictive worth.

Content validity has two requirements that must be met for a grading system to be considered adequate. The system must focus evaluators on student writing as writing, and it must focus evaluators on applying the appropriate criteria. CHE meets these needs in two ways. First, the criteria themselves are arrived at through thorough discussion of an assignment and collaborative

---

1 This assumes that the student’s health and devotion to school, and other factors, remain the same.
development of the key features specific to that assignment. These features then become integral parts of the evaluation rubric that is developed immediately prior to grading sessions, which has the effect of focusing us on applying the proper criteria. Second, the post-session comparisons and discussions of the independently assigned grades serve to ensure that we are applying the criteria, and that we are focusing on student writing rather than on personal experiential responses.

Construct validity, too, has two requirements that an evaluation system must fulfill if it is to be adequate. The system must provide some way to focus evaluators on the key evaluative features, and the same assignment must be used repeatedly and the papers for that assignment evaluated repeatedly using the same evaluation system. As the discussion earlier in this chapter shows, CHE focuses us on the key features of a specific assignment by being a part of assignment development and by requiring us to develop the rubric immediately prior to grading sessions. The same assignment is used repeatedly by virtue of the facts that Miller and I give our separate classes the same assignment sheet and discuss the same problems with our classes, and the evaluation system is used repeatedly to evaluate that assignment by virtue of the fact that papers from the separate classes are independently evaluated by both of us using that system.
In essence, CHE is a valid system that arrives at valid grades because it relies on the same collaborative methodology for evaluating performance that is used for generating the knowledge necessary to perform. In other words, the conversations that Miller and I have as members of the community of writing experts, and the conversations we have with our students in our classes, are mirrored in the conversations they have with us in their papers and that we have in evaluating those papers.

Reliability

As noted earlier, reliability is simply a matter of consistency, and an evaluation system is reliable if it consistently results in the same grade being given to a paper after multiple readings. Under this standard, CHE can be considered reliable if Miller and I consistently give either the same grades or grades no more than one step apart to at least seventy percent of the papers we evaluate (Huot 202). Below is a table that sets out the evaluation data compiled from the two writing courses I taught Spring semester of 1991, the first semester we used CHE. As Table 1 shows, after the first set of papers, Miller and I always achieved consistent scores in excess of the seventy percent minimum required to establish an evaluation system as reliable. As roughly equal figures were achieved in the second and third semesters of use, it is evident that CHE is a reliable form of evaluation and grading.
Table 1. CHE Reliability Figures for Spring 1991

<table>
<thead>
<tr>
<th>Paper #</th>
<th>Paper Count</th>
<th>Same Grade</th>
<th>Diff. Grade</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
<td>31</td>
<td>15</td>
<td>67%</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>35</td>
<td>9</td>
<td>79%</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>36</td>
<td>5</td>
<td>88%</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>36</td>
<td>6</td>
<td>86%</td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>41</td>
<td>6</td>
<td>87%</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>33</td>
<td>9</td>
<td>78%</td>
</tr>
<tr>
<td>7</td>
<td>41</td>
<td>32</td>
<td>9</td>
<td>78%</td>
</tr>
</tbody>
</table>

**Topic Bias**

That CHE provides us with a mechanism to control topic bias can best be illustrated with an anecdote about a paper from the second group of papers we evaluated using CHE (i.e. the group in which we first achieved acceptable reliability). Not surprisingly, it was one of the most wildly discrepant papers we have ever had.

The assignment was for students to write about places that had some special significance for them. The criteria for evaluation were: (1) the use of vivid, sensuous details to describe the place, (2) a statement of the place’s significance, and (3) an explanation of why the place had its significance to the writer. One student wrote about a tree that he had been climbing and sitting in since he was a small boy. The description of the tree was general rather than
specific, the statement of significance was a short sentence, and the explanation of significance was rather trite and boring. All-in-all, the paper was clearly a C level piece of writing.

However, one of us had had exactly such a tree in his life—a tree that was climbed and loved, that provided refuge from the family and a place to contemplate, a tree that bore his initials. To this evaluator, the paper was an A with starbursts.

In the post-session discussion, the tree paper was thoroughly examined, and the latter evaluator arrived at the conclusion that he was responding more powerfully to his own memories than to the student's skills as a writer. Discounting his emotional response to his experience, he concluded that the paper was a C level paper, and that was the final grade given to the paper.

As this anecdote shows, CHE has a two-step mechanism to control topic bias. First, an idiosyncratic response is likely to trigger a significantly higher or lower grade than is merited by the student's demonstrated writing ability, a grade that will not be duplicated within acceptable limits by the other evaluator. Second, that idiosyncrasy will be revealed in the post-session discussion, and the emotional response can then be incorporated into a grade rather than permitted to dominate a grade.
Summary

Obviously, the foregoing discussion is but a brief glimpse of CHE and its larger collaborative teaching framework. Other elements of our collaborative approach to teaching, such as class activities and demonstrations of CHE to our classes, have an impact on our evaluation system and grades, and CHE itself contributes to our efforts to diagnose writing problems and suggest solutions. Thus, it is apparent that this thesis suffers from the same analytical disabilities that plague any partial explanation of a whole.

Nevertheless, the problems of validity, reliability, and topic bias, the theoretical underpinnings of collaboration and holisticism, and the methodologies of CHE have all been discussed and their interworkings elaborated. Hopefully, some component of that discussion will be beneficial to the reader and the reader's students.
CONCLUSION

As shown in the first chapter, evaluating and grading student papers is a far from simple task. Among other problems, teachers face the difficulties of giving valid, reliable grades, and avoiding topic bias. Although no system yet developed perfectly solves these (or other) problems, holistic systems offer an approach based on sound theoretical principles and highly practical methodologies.

As set out in Chapter Two, holistic evaluation derives theoretical support from two basic principles: (1) knowledge is created, maintained, and transferred through collaborative acts, and (2) adequate evaluation of a text demands that it be viewed as a whole, not simply as a collection of parts. In essence, holistic systems seek to mirror both the collaborative processes of creating texts and the holistic nature of final products.

Methodologically, holistic evaluation relies on a set of seven basic features designed to implement its theoretical foundations. The controlled reading environment, impressionistic scoring, use of multiple readers, etc., are pragmatic instantiations of the theoretical principles of collaboration and holism.

Collaborative Holistic Evaluation (CHE), as created by Walden Miller and myself, is one such holistic system, developed strictly for use in the composition classroom. Based on the anti-reductionist and collaborative foundations common to all holistic
systems, CHE provides us with a means to assess student writing so as to arrive at fair and accurate grades.

However, much remains to be explored, and more questions need to be answered, before CHE can be considered a thoroughly documented, thoroughly examined system for evaluating student papers. One important area for study, because of its potential as a source of bias, is the evaluators. Specifically, how important to reliability are the relationships among evaluators? Is CHE usable only by colleagues who are friends, or can it be used by those on less than the best of terms? My experiences with Miller, who is both a respected professional colleague and a close personal friend, lead me to the tentative hypothesis that amicable relationships between evaluators are necessary for two reasons. First, mutual respect and liking encourage honest communication at all the stages of evaluation, which makes it easier either to resolve disagreements or to respectfully disagree. Second, a solid foundation of friendship eliminates ego-battles fought over the ground of students' papers—attention is focused on students' texts as texts rather than on texts as turf to be won or lost.

Another important area for study is the impact of CHE on students. Since students are the beneficiaries (or victims) of evaluation systems, the impact of those systems on students can hardly be over-emphasized. Among other important questions that
need answering are: Does CHE make instructors better evaluators? Does CHE make its users better teachers, and if so, how?

Again, my personal feeling is that CHE makes me better both as an evaluator and as a teacher. When Miller and I discuss an assignment, I cannot rely on the comfortable laziness of mentally saying, "I know what I mean." Instead, I must be able to articulate what I mean; I must be able to communicate my understanding of an assignment to another professional. When Miller and I discuss a student's paper, I cannot simply judge that paper; I must also support my judgment with references to the text. And those conversations with Miller keep me honest with my students and myself.

There are more areas to be studied and questions to be answered. However, with its firm theoretical underpinnings and sound methodological processes, CHE is a valid and valuable tool for arriving at fair and accurate grades.
WORKS CITED


Freedman, Sarah W. "Why Do Teachers Give the Grades They Do?" College Composition and Communication 30 (1979): 161-64.


WORKS CONSULTED


ACKNOWLEDGMENTS

The list of persons I should acknowledge is far too lengthy to be produced in its entirety, so I will restrict myself to four individuals, each of whom has provided me with the sort of encouragement, ideas, and support that made graduate school, and this thesis, possible.

David Roberts—Coach—who has pretty much allowed me to do things my way while helping me to learn other and usually better ways.

Linda and Aubrey Galyon, who both had faith in me and knew me well enough to know that one day I would return and finish this. It took thirteen years, but they were right.

And Walden Miller, who knows exactly what I mean when I say I feel old, and without whose work and support this thesis literally could not have been written.

Thank you all.